UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

NOTICE OF ALLOWANCE AND FEE(S) DUE

7590 11/30/2004

J.C. Patents, Inc.
Suite 114
1340 Reynolds Ave.
Irvine, CA 92614

EXAMINER	
FREJD, RUSSELL WARREN	
ART UNIT	PAPER NUMBER
2128	

DATE MAILED: 11/30/2004

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/920,034	08/01/2001	Chen-Fu Chien	JCLA6835	5484

TITLE OF INVENTION: OVERLAY ERROR MODEL, SAMPLING STRATEGY AND ASSOCIATED EQUIPMENT FOR IMPLEMENTATION

APPLN. TYPE	SMALL ENTITY	ISSUE FEE	PUBLICATION FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	NO	$1370	$300	$1670	02/28/2005

THE APPLICATION IDENTIFIED ABOVE HAS BEEN EXAMINED AND IS ALLOWED FOR ISSUANCE AS A PATENT. PROSECUTION ON THE MERITS IS CLOSED. THIS NOTICE OF ALLOWANCE IS NOT A GRANT OF PATENT RIGHTS. THIS APPLICATION IS SUBJECT TO WITHDRAWAL FROM ISSUE AT THE INITIATIVE OF THE OFFICE OR UPON PETITION BY THE APPLICANT. SEE 37 CFR 1.313 AND MPEP 1308.

THE ISSUE FEE AND PUBLICATION FEE (IF REQUIRED) MUST BE PAID WITHIN THREE MONTHS FROM THE MAILING DATE OF THIS NOTICE OR THIS APPLICATION SHALL BE REGARDED AS ABANDONED. THIS STATUTORY PERIOD CANNOT BE EXTENDED. SEE 35 U.S.C. 151. THE ISSUE FEE DUE INDICATED ABOVE REFLECTS A CREDIT FOR ANY PREVIOUSLY PAID ISSUE FEE APPLIED IN THIS APPLICATION. THE PTOL-85B (OR AN EQUIVALENT) MUST BE RETURNED WITHIN THIS PERIOD EVEN IF NO FEE IS DUE OR THE APPLICATION WILL BE REGARDED AS ABANDONED.

HOW TO REPLY TO THIS NOTICE:

I. Review the SMALL ENTITY status shown above.

If the SMALL ENTITY is shown as YES, verify your current SMALL ENTITY status:

A. If the status is the same, pay the TOTAL FEE(S) DUE shown above.

B. If the status above is to be removed, check box 5b on Part B - Fee(s) Transmittal and pay the PUBLICATION FEE (if required) and twice the amount of the ISSUE FEE shown above, or

If the SMALL ENTITY is shown as NO:

A. Pay TOTAL FEE(S) DUE shown above, or

B. If applicant claimed SMALL ENTITY status before, or is now claiming SMALL ENTITY status, check box 5a on Part B - Fee(s) Transmittal and pay the PUBLICATION FEE (if required) and 1/2 the ISSUE FEE shown above.

II. PART B - FEE(S) TRANSMITTAL should be completed and returned to the United States Patent and Trademark Office (USPTO) with your ISSUE FEE and PUBLICATION FEE (if required). Even if the fee(s) have already been paid, Part B - Fee(s) Transmittal should be completed and returned. If you are charging the fee(s) to your deposit account, section "4b" of Part B - Fee(s) Transmittal should be completed and an extra copy of the form should be submitted.

III. All communications regarding this application must give the application number. Please direct all communications prior to issuance to Mail Stop ISSUE FEE unless advised to the contrary.

IMPORTANT REMINDER: Utility patents issuing on applications filed on or after Dec. 12, 1980 may require payment of maintenance fees. It is patentee's responsibility to ensure timely payment of maintenance fees when due.

PTOL-85 (Rev. 11/04) Approved for use through 04/30/2007.

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), to: Mail **Mail Stop ISSUE FEE**
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
or Fax (703) 746-4000

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

7590 11/30/2004

J.C. Patents, Inc.
Suite 114
1340 Reynolds Ave.
Irvine, CA 92614

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission
I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being facsimile transmitted to the USPTO (703) 746-4000, on the date indicated below.

(Depositor's name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/920,034	08/01/2001	Chen-Fu Chien	JCLA6835	5484

TITLE OF INVENTION: OVERLAY ERROR MODEL, SAMPLING STRATEGY AND ASSOCIATED EQUIPMENT FOR IMPLEMENTATION

APPLN. TYPE	SMALL ENTITY	ISSUE FEE	PUBLICATION FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	NO	$1370	$300	$1670	02/28/2005

EXAMINER	ART UNIT	CLASS-SUBCLASS
FREJD, RUSSELL WARREN	2128	703-002000

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).
- ☐ Change of correspondence address (or Change of Correspondence Address form PTO/SB/122) attached.
- ☐ "Fee Address" indication (or "Fee Address" Indication form PTO/SB/47; Rev 03-02 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list
(1) the names of up to 3 registered patent attorneys or agents OR, alternatively,
(2) the name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.
1 _____
2 _____
3 _____

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document has been filed for recordation as set forth in 37 CFR 3.11. Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE (B) RESIDENCE: (CITY and STATE OR COUNTRY)

Please check the appropriate assignee category or categories (will not be printed on the patent) : ☐ Individual ☐ Corporation or other private group entity ☐ Government

4a. The following fee(s) are enclosed:
- ☐ Issue Fee
- ☐ Publication Fee (No small entity discount permitted)
- ☐ Advance Order - # of Copies _____

4b. Payment of Fee(s):
- ☐ A check in the amount of the fee(s) is enclosed.
- ☐ Payment by credit card. Form PTO-2038 is attached.
- ☐ The Director is hereby authorized by charge the required fee(s), or credit any overpayment, to Deposit Account Number _____ (enclose an extra copy of this form).

5. **Change in Entity Status** (from status indicated above)
- ☐ a. Applicant claims SMALL ENTITY status. See 37 CFR 1.27.
- ☐ b. Applicant is no longer claiming SMALL ENTITY status. See 37 CFR 1.27(g)(2).

The Director of the USPTO is requested to apply the Issue Fee and Publication Fee (if any) or to re-apply any previously paid issue fee to the application identified above.
NOTE: The Issue Fee and Publication Fee (if required) will not be accepted from anyone other than the applicant; a registered attorney or agent; or the assignee or other party in interest as shown by the records of the United States Patent and Trademark Office.

Authorized Signature _____ Date _____

Typed or printed name _____ Registration No. _____

Notice of Allowability	Application No.	Applicant(s)	
	09/920,034	CHIEN ET AL.	
	Examiner	Art Unit	
	Russell Frejd	2128	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☒ This communication is responsive to *applicant's filing on 1-August-2001*.

2. ☒ The allowed claim(s) is/are *1-5*.

3. ☒ The drawings filed on *01 August 2001* are accepted by the Examiner.

4. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
   a) ☐ All b) ☐ Some* c) ☐ None of the:
      1. ☐ Certified copies of the priority documents have been received.
      2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
      3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____.

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

5. ☐ A SUBSTITUTE OATH OR DECLARATION must be submitted. Note the attached EXAMINER'S AMENDMENT or NOTICE OF INFORMAL PATENT APPLICATION (PTO-152) which gives reason(s) why the oath or declaration is deficient.

6. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   (a) ☐ including changes required by the Notice of Draftsperson's Patent Drawing Review (PTO-948) attached
      1) ☐ hereto or 2) ☐ to Paper No./Mail Date _____.
   (b) ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____.

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

7. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☒ Notice of References Cited (PTO-892)
2. ☐ Notice of Draftperson's Patent Drawing Review (PTO-948)
3. ☐ Information Disclosure Statements (PTO-1449 or PTO/SB/08), Paper No./Mail Date _____
4. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material
5. ☐ Notice of Informal Patent Application (PTO-152)
6. ☐ Interview Summary (PTO-413), Paper No./Mail Date _____ .
7. ☒ Examiner's Amendment/Comment
8. ☒ Examiner's Statement of Reasons for Allowance
9. ☐ Other _____.

Russell Frejd
RUSSELL FREJD
PRIMARY EXAMINER

Allowance of Application # 09/920,034

1. The following communication is in response to applicant's filing on 1-August-2001.

Examiner's Amendment

2. An Examiner's Amendment to the record appears below. Should the changes and/or additions be unacceptable to applicant, an amendment may be filed as provided by 37 C.F.R. 1.312. To ensure consideration of such an amendment, it **MUST** be submitted no later than the payment of the Issue Fee. Authorization for this Examiner's Amendment was given by Jiawei Huang (Reg. No. 43,330) on November 10 and 18, 2004.

3. In the Claims:

In Claim 1:

line 3 Add --the-- between "determining" and "number".

line 8 Add --the-- between "using" and "measured".

lines 11-12 Delete "(or R-square represents degree of assessed variance explained by the model)" and Add --, where R-square represents the degree of assessed variance explained by the model--.

line 28 Delete "the same" and Add –a--.

Add --the-- between "that" and "number".

line 29 Delete "to a tolerable range" and change "yield of semiconductor" to --the yield of the semiconductor--.

line 31 Change "errors" to --error-- and "follow" to --follows--.

In Claim 3:

line 2 Change "intra-field" to --inter-field--.

In Claim 6:

Cancel claim 6.

Reasons for Allowance

4. The following is an Examiner's Statement of Reasons for the indication of allowable subject matter. The instant application is directed to a non-obvious improvement over the information described in the article authored by Chien et al., entitled "Sampling Strategy and Model to Measure and Compensate the Overlay Errors", *Metrology, Inspection, and Process Control for Microlithography XV*, Proceedings-of-the-SPIE-The-International-Society-for-Optical-Engineering, Vol. 4344, 26 February-1 March, 2001, p. 245-256.

The improvement comprises a strategic sampling procedure for measuring overlay errors in the manufacture of semiconductor wafers, wherein integrative overlay error models, including parameters for inter-field translation, magnification, rotation, expansion and non-positive crossing, as well as intra-field translation, magnification and rotation, are used to provide a model for assessing and minimizing overlay errors and providing an optimal sampling strategy.

This patentable distinction is included in independent claim no. 1. The art of record, either individually or in combination, fails to teach, suggest, or render obvious the useful, concrete and tangible <integrative overlay model comprised of equations 23 and 24, the corresponding model parameters and the sampling strategy,> having the corresponding structure which is disclosed in the specification and equivalents thereof (at least at page 16, line

2 through page 29, line 22, and Figures 1-21). In view of the foregoing, the claims of the present application are found to be patentable over the prior art.

Response Guidelines

5. Any comments considered necessary by applicant **MUST** be submitted no later than the payment of the Issue Fee and, to avoid processing delays, should preferably accompany the Issue Fee. Such submissions should clearly be labeled "Comments on Statement of Reasons for Allowance".

6. Any response to the Examiner in regard to this allowance should be

directed to: Russell Frejd, telephone number (571) 272-3779, Monday-Friday from 0530 to 1400 ET, **or** the examiner's supervisor, Jean Homere, telephone number (571) 272-3780.

mailed to: Commissioner of Patents and Trademarks
Washington, D.C. 20231

or faxed to: (571) 273-3779

Hand-delivered responses should be brought to 220 South 20th Street, Crystal Plaza Two, Lobby, Room 1B03, Arlington. VA., 22202.

Date: 12-November-2004

Russell Frejd

RUSSELL FREJD
PRIMARY EXAMINER

Notice of References Cited	Application/Control No. 09/920,034	Applicant(s)/Patent Under Reexamination CHIEN ET AL.	
	Examiner Russell Frejd	Art Unit 2128	Page 1 of 2

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-6,556,959	04-2003	Miller et al.	703/2
	B	US-6,535,774	03-2003	Bode et al.	700/109
	C	US-6,442,496	08-2002	Pasadyn et al.	702/83
	D	US-5,805,866	09-1998	Magome et al.	716/19
	E	US-5,502,311	03-1996	Imai et al.	250/548
	F	US-5,498,501	03-1996	Shimoda et al.	430/22
	G	US-5,448,333	09-1995	Iwamoto et al.	355/53
	H	US-4,918,320	04-1990	Hamasaki et al.	250/548
	I	US-4,833,621	05-1989	Umatate, Toshikazu	716/21
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	BULLER et al., Manufacturing Issues Related to RTP Induced Overlay Errors in a Global Alignment Stepper Technology, IEEE Transactions on Semiconductor Manufacturing, Vol. 9, No. 1, February 1996, p. 108-114.
	V	HEBB et al., The Effect of Patterns on Thermal Stress During Rapid Thermal Processing of Silicon Wafers, IEEE Transactions on Semiconductor Manufacturing, Vol. 11, No. 1, February 1998, p. 99-107.
	W	PREIL et al., A New Approach to Correlating Overlay and Yield, SPIE Conference on Metrology, Inspection, and Process Control, for Microlithography XIII, Vol. 3677, March 1999, p. 208-16.
	X	SHAMOUN et al., Assessment of Thermal Loading-Induced Distortions in Optical Photomasks Due to e-Beam Multipass Patterning, 42nd Int. Con. on Electron, Ion, and Photon Beam Tech/Nanofabrica-tion, American Vacuum Society, Nov/Dec 1998, p. 3558-62.

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Notice of References Cited	Application/Control No. 09/920,034	Applicant(s)/Patent Under Reexamination CHIEN ET AL.	
	Examiner Russell Frejd	Art Unit 2128	Page 2 of 2

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-			
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	GOODWIN et al., Characterizing Overlay Registration of Concentric 5X and 1X Stepper Exposure Fields Using Interfield Data, SPIE Conference on Metrology, Inspection, and Process Control for Microlithography XI, Vol. 3050, March 1997, p. 407-17.
	V	CHIEN et al., Sampling Strategy and Model to Measure and Compensate the Overlay Errors, SPIE Conference on Metrology, Inspection, and Process Control for Microlithography XV, Vol. 4344, Feb/Mar 2001, p. 245-56.
	W	HONG et al., Interfield Sampling Method Dependency of Overlay and Global Alignment, SPIE Conference on Metrology, Inspection, and Process Control for Microlithography XIV, Vol. 3998, Feb/Mar 2000, p. 856-62.
	X	ARNOLD, Overlay Simulator for Wafer Steppers, SPIE Vol. 922, Optical/Laser Microlithography, March 1988, p. 94-105.

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Manufacturing Issues Related to RTP Induced Overlay Errors in a Global Alignment Stepper Technology

James F. Buller, *Member, IEEE*, M. M. Farahani, and Shyam Garg

Abstract— **The effect of rapid thermal processing on wafer distortion and overlay accuracy in global alignment photolithography in the fabrication of 0.85 μm CMOS Flash EPROM integrated circuits was studied. Both rapid thermal process parameters and system design (single and multi-lamp processors) were evaluated for their effect on overlay accuracy. It was found that a rapid thermal process (following contact etch and ion implantation) at set temperatures greater than or equal to 950°C resulted in interconnect metallization-to-contact overlay errors in excess of 1.0 μm across the wafer, which led to a 20% functional circuit yield loss. In the case of the single lamp processor, this misalignment was attributed to wafer distortion due to the temperature overshoot during the ramp step, which subsequently resulted in an across wafer temperature range of greater than 120°C. This temperature overshoot and nonuniformity was eliminated by reducing the ramp rate below 100°C/s. This ramp rate reduction, however, decreased the system wafer throughput, and required optimization to eliminate the overlay errors and minimize the effect on throughput. In this study, a 60°C/s ramp rate was found to be optimum. For the multi-lamp RTP system, the metal-to-contact overlay error was not observed. This was believed to be due to the design of the heating mechanism in the multi-lamp processor, which did not produce the large wafer temperature overshoot and nonuniformity that was observed in the single lamp processor.**

I. Introduction

SHRINKING device geometries and the demand for high wafer throughput place a great demand on semiconductor process integration. This has resulted in rapid thermal processing (RTP) becoming an essential part of submicron semiconductor Integrated Circuit (IC) fabrication. Some potential applications for RTP in submicron semiconductor processing include: oxide growth, Chemical Vapor Deposition (CVD), silicidation, and ion implant dopant activation and implant damage repair with minimum lateral dopant diffusion [1]–[10]. Due to the difficulties in achieving excellent temperature and across wafer uniformity control, RTP has been somewhat slow to gain wide acceptance in semiconductor processing. For ion implant activation (see Table I), a nonuniform temperature distribution across the wafer during the RTP can result in permanent wafer distortion [11]. This distortion can result in overlay (alignment) errors at a subsequent masking step [12], particularly in the case of global alignment stepper photolithography. Gross wafer distortion prior to a critical masking step such as interconnect metallization can lead to significant functional circuit loss at test.

TABLE I
PROCESS FLOW FOR EXPERIMENTAL TECHNOLOGY

1. Contact Etch
2. Implant
3. RTP (950°C)
4. Barrier Metal Deposition and RTP (600°C)
5. W Plug
6. Al/Si/Cu Deposition
7. Metal Masking

Typically in global alignment lithography, two alignment targets are placed on the wafer. During alignment, the stepper measures the distance between the two targets and calculates a wafer "scaling". This scaling is used by the stepper to account for changes in the separation of the alignment targets due to previous processing. Normally this scaling is calculated for one direction (for example the x-direction), and then assumed to be the same for the perpendicular direction. This allows the semiconductor manufacturer to use only two alignment targets, thereby minimizing wafer real estate not utilized by product die. Once the scaling has been calculated (along with rotation, etc.) the stepper then sequentially aligns the entire wafer without any additional measurements. If the scaling between the alignment targets is not representative of wafer distortion everywhere, then misalignments (overlay errors) above and beyond the stepper overlay accuracy capability could occur. In the case of site-by-site alignment photolithography, this problem is not as severe, since alignment targets are identified by the stepper for each exposure field on the wafer. However, the wafer throughput for the global alignment stepper is significantly higher than that of a comparable site-by-site alignment system.

Rapid thermal processing and processor design and capabilities are rapidly becoming of great importance to semiconductor manufacturing engineers. Therefore, the objective of this work was to investigate the effect of rapid thermal processing on interconnect metallization masking overlay accuracy in a 0.85 μm CMOS Flash EPROM process based on global alignment. The RTP temperature and uniformity, ramp rate, and the rapid thermal processor design were evaluated for their effects on overlay accuracy.

Manuscript received Octber 20, 1994; revised April 20, 1995, and May 16, 1995.

The authors are with Advanced Micro Devices, Austin, TX 78741 USA.

Publisher Item Identifier S 0894-6507(96)01138-4.

TABLE II
EXPERIMENTAL FACTORS AND RANGE

Factor	Range
Temperature	900-1000°C
Ramp Rate	20-100°C/Second
RTP System	Single Lamp
	Multi-Lamp

II. EXPERIMENTAL

A 0.85 μm, a 1 Mbit Flash EPROM device was used as the vehicle to investigate the effects of RTP on photolithography overlay. Device design and process thermal budget constraints in an experimental technology required that a rapid thermal process be performed following contact etch and ion implantation. Table I shows the process flow between contact etch and interconnect metallization masking where the excessive overlay error was observed in an experimental technology. A control technology without a high temperature (> 600°C) RTP step did not exhibit the metal-to-contact overlay error problem. Experiments were conducted to isolate the source of the metal-to-contact misalignment in the experimental technology containing the 950°C RTP step. These experiments included aligning the metal mask pattern to the wafer immediately following each of steps one through six individually. This series of experiments clearly showed that the 950°C RTP step caused metal-to-contact overlay errors greater than the stepper overlay accuracy.

Based on the results of the above experiments, a detailed evaluation of the RTP and its effect on lithography overlay was conducted. Table II shows the range of temperature, ramp rate, and RTP system that were investigated. The single lamp system employed a single tubular high intensity lamp with parabolic reflector and quartz diffuser plate to radiantly heat the wafer only from the front-side. The multi-lamp design utilized two banks of high intensity tungsten halogen lamps to heat the wafer from both the front and back-sides of the wafer. The temperature in the single-lamp system was controlled by lamp power, which was calibrated from temperature measurements performed with a thermocouple attached to the center of a silicon wafer. A calibration table was then generated from these measurements, which the system used to adjust lamp power to obtain the desired temperature. Therefore, the temperature control in the single-lamp processor was independent of the wafer backside emissivity. Output from a backside optical pyrometer aimed at the wafer center was used to control power to the lamps for the multi-lamp processor, making this temperature sensing scheme dependent on wafer backside emissivity.

The first experiment investigated the effect of process temperature (at steady state) of the RTP step on the wafer distortion and metal-to-contact alignment. Three, 48 wafer 1 Mbit Flash EPROM lots were evenly split between RTP process temperatures of 900, 950, and 1000°C in the single lamp system. The ramp rate in all of these cases was 100°C/s. Both pre and post-RTP wafer deflection measurements were performed on ten wafers per temperature split per wafer lot using a single point Ionics Strain Gauge. This deflection measurement yielded a relative measure of the degree of wafer distortion due to the RTP processing, and demonstrated whether the distortion was concave or convex. Alignment accuracy between the metal interconnect photoresist pattern and the etched contact pattern was obtained on a Nikon inspection station where alignment verniers were measured for eleven positions per wafer on ten wafers for each RTP temperature treatment on all three product lots. The 3σ overlay accuracy of the ASML PAS2 500/40 i-line stepper was 0.3 μm.

The second series of experiments examined the effects of temperature ramp rate on the temperature control and uniformity in the single lamp rapid thermal processor. A SensArray [13] temperature measurement system was used to measure and wafer map the temperature distribution as a function of temperature ramp rate in the single lamp system. Temperature responses for ramp rates between 20 and 100°C/s were investigated and compared. In addition, a two-step ramp rate process, whereby the temperature was ramped to 900°C at 60°C/s and then to 950°C at 20°C/s was also evaluated for temperature control and uniformity.

A special *product* wafer with 17 attached thermocouples was constructed for these experiments. This wafer had all of the films and topography of a product wafer prior to the implant RTP step. The thermocouples were distributed at various radial positions across the wafer as illustrated in Fig. 1. A product wafer was chosen to simulate the wafer backside emissivity used for temperature control in the multi-lamp processor, however, nearly identical results for wafer distortion and wafer temperature variation were obtained with bare silicon wafers. As discussed earlier, the single-lamp processor was modified by the manufacturer and AMD to make the temperature sensing independent of backside emissivity.

Finally, an evaluation of the temperature response of a multi-lamp RTP system was characterized as a function of ramp rate using the SensArray. The results were then compared to those obtained with the single lamp processor.

III. RESULTS AND DISCUSSION

Fig. 2 shows the results of the Ionics wafer deflection measurements pre and post-RTP measured on 30 wafers at each of the RTP temperatures. The temperature ramp rate was 100°C/s for the three temperature treatments of this experiment. The wafer deflection data showed that at the RTP temperature of 900°C, the wafer delta deflection (post RTP-pre RTP) was small, concave, and very tightly distributed. For 950°C, the variability in wafer delta deflection increased, and several of the wafer deflections were convex. At 1000°C, the majority of wafers had a convex wafer deflection and several of the wafers had large wafer delta deflections of thousands of μin.

A summary of the metal-to-contact misalignment data for two typically affected wafer locations for rapid thermal processing in the single-lamp system is shown in Table III. In the case of the control technology (no RTP > 600°C) no overlay errors greater than 0.3 μm were observed at interconnect metallization. Both the 950 and 1000°C RTP induced metal-

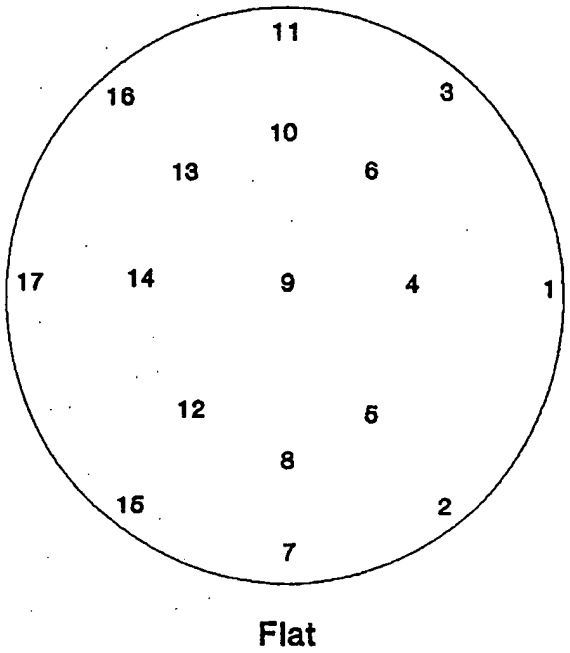


Fig. 1. Radial distribution of 17 thermocouples attached to SensArray wafer.

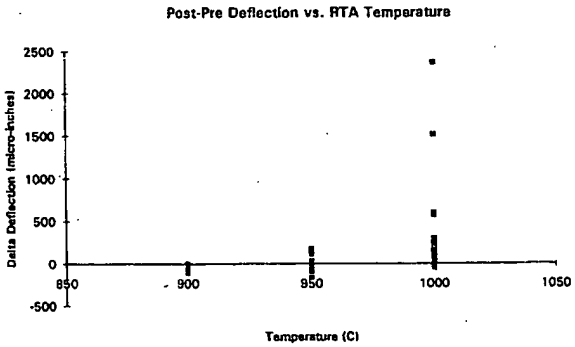


Fig. 2. Post-RTP-Pre-RTP wafer deflection versus RTP temperature measured on wafers using an ionics strain gauge.

to-contact misalignments in excess of the 0.3 μm ASM stepper overlay accuracy (Table III). The 1000°C RTP produced much larger misalignment errors with larger variability, as well as a significantly higher percentage of wafers with greater than 0.3 μm metal-to-contact misalignment (%Fail). No misalignment errors greater than 0.3 μm were observed on wafers processed at 900°C. These data correlated well with the wafer deflection data. The above misalignment data also agreed qualitatively with that obtained by Feil *et al.* in [12]. Feil *et al.* found that reflow in a front-side only rapid thermal processor at temperatures greater than 1000°C produced subsequent pattern overlay errors in excess of 1.0 μm. In the present work, gross pattern misalignment was observed for an RTP set temperature as low as "950°C". Fig. 3 shows a typical wafer map of the metal-to-contact misalignment observed for an RTP set temperature of 950°C with a ramp rate of 100°C/s

TABLE III
SUMMARY OF MISALIGNMENT DATA FOR SINGLE-LAMP PROCESSOR

Process	Location	Misalignment (μm)	3σ (μm)	Range (μm)	%Fail (%)
900°C	1	0.119	0.285	0.30	0
	2	0.094	0.261	0.30	0
950°C	1	0.183	0.402	0.40	16.7
	2	0.170	0.357	0.40	6.7
1000°C	1	0.412	0.675	0.90	60.0
	2	0.326	0.624	0.70	53.3

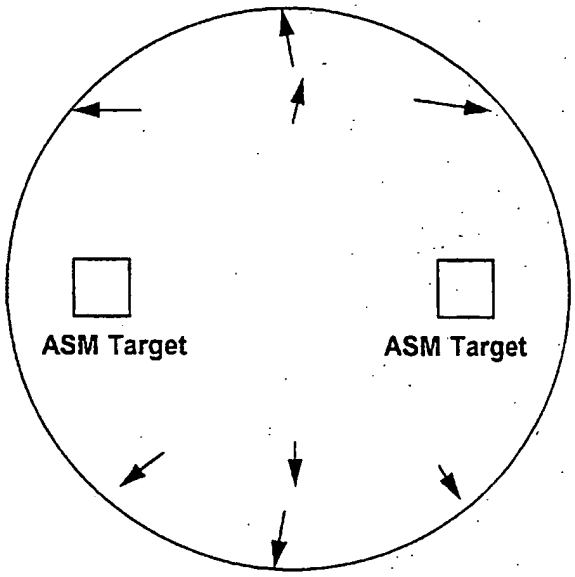


Fig. 3. Typical wafer map of metal-to-contact misalignment for the 950°C RTP in the single lamp processor with a ramp rate of 100°C/s.

in the single lamp processor. The y-direction misalignment scaled linearly through the wafer center perpendicular to the major flat. Also, both x- and y-direction misalignment was observed near the wafer perimeter. Interconnect metallization-to-contact misalignment exceeded 1.0 μm in extreme cases. This resulted in metal lines completely missing contacts, thereby causing integrated circuit functional failures. Fig. 4 shows a cross-section SEM micrograph of severe metal-to-contact misalignment observed within a 1 Mbit EPROM near the wafer perimeter. In a perfectly aligned device, the metal 1 lines in Fig. 4 would be centered directly over the tungsten plugged contacts.

Fig. 5 shows a temperature profile for the "950°C" (ramp rate = 100°C/s) RTP in the single lamp (front-side illumination) processor obtained with the 17 thermocouple SensArray system. The individual traces in this figure represent the temperature response of the 17 thermocouples. The temperature at certain locations near the wafer perimeter (~5–10 mm from the wafer edge) overshot the 950°C set temperature during an approximately 5 second transient time period, where wafer temperatures exceeded 1020°C, an overshoot of more than

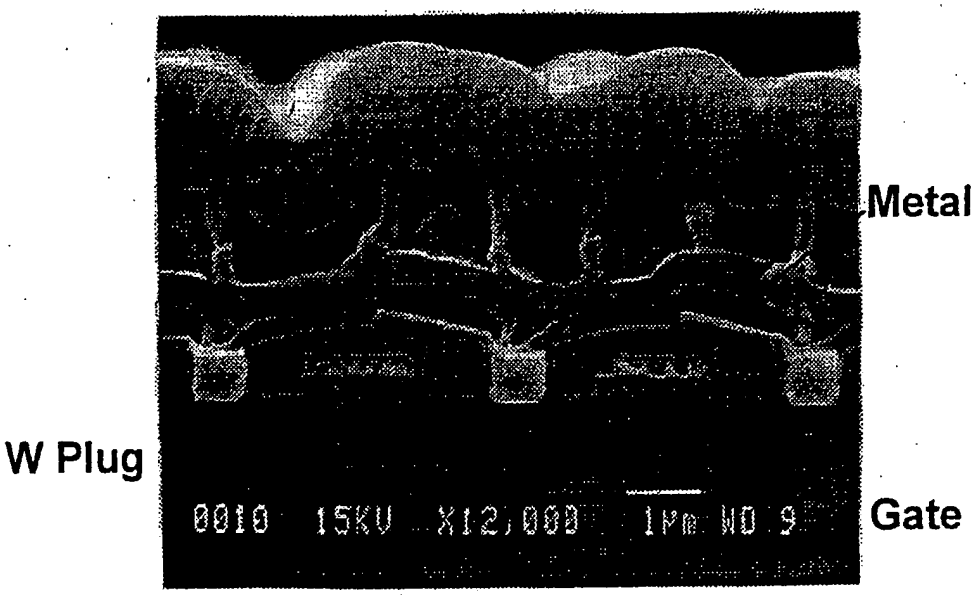


Fig. 4. Cross-section SEM micrograph of metal-to-contact misalignment within a 1 Mbit EPROM.

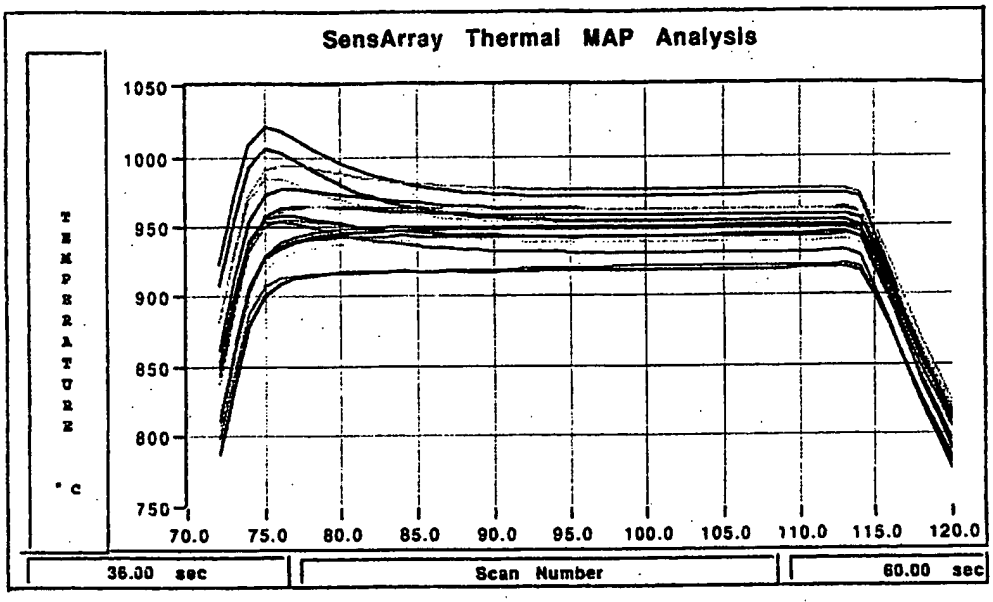


Fig. 5. Temperature profile for "950°C" RTP (ramp rate = 100°C/s) in single lamp rapid thermal processor measured with the 17 thermocouple SensArray temperature monitor system.

70°C. The wafer center was at a much cooler temperature of approximately 900°C. This resulted in an across wafer temperature range of more than 120°C. Thus, even though a set (mean steady state) temperature of 950°C was obtained, transient overshoot temperatures in excess of 1020°C were realized in the single lamp RTP system with a temperature ramp rate of 100°C/s. This extreme temperature overshoot near the wafer perimeter in the single lamp processor caused the wafer deflection and subsequent metal-to-contact alignment errors with this process technology.

In order to understand the temperature overshoot in the single lamp processor, experiments with ramp rate were conducted. SensArray measurements showed that the temperature overshoot (and across wafer temperature range) was reduced by approximately 30% when the ramp rate was changed from 100°C/s to 60°C/s. Instead of wafer transient overshoot temperatures in excess of 1020°C, the highest temperature with the reduced ramp rate was 1000 °C, an overshoot of approximately 50°C. This overshoot was further reduced by a two-step ramp rate process, where the temperature was ramped to 900°C at

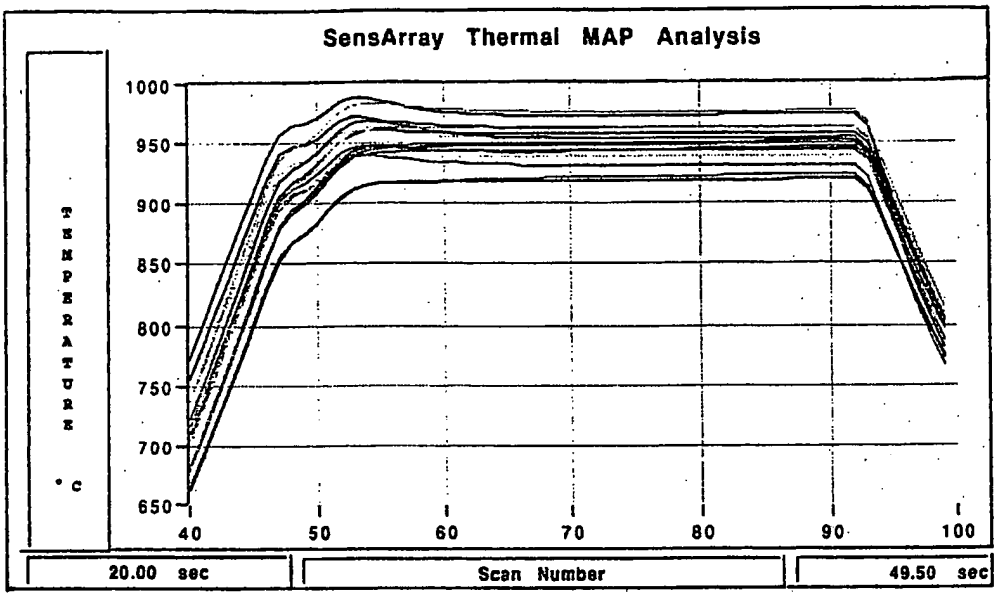


Fig. 6. Temperature profile for "950°C" two-step ramp rate RTP in the single lamp rapid thermal processor measured with the 17 thermocouple SensArray temperature monitor system.

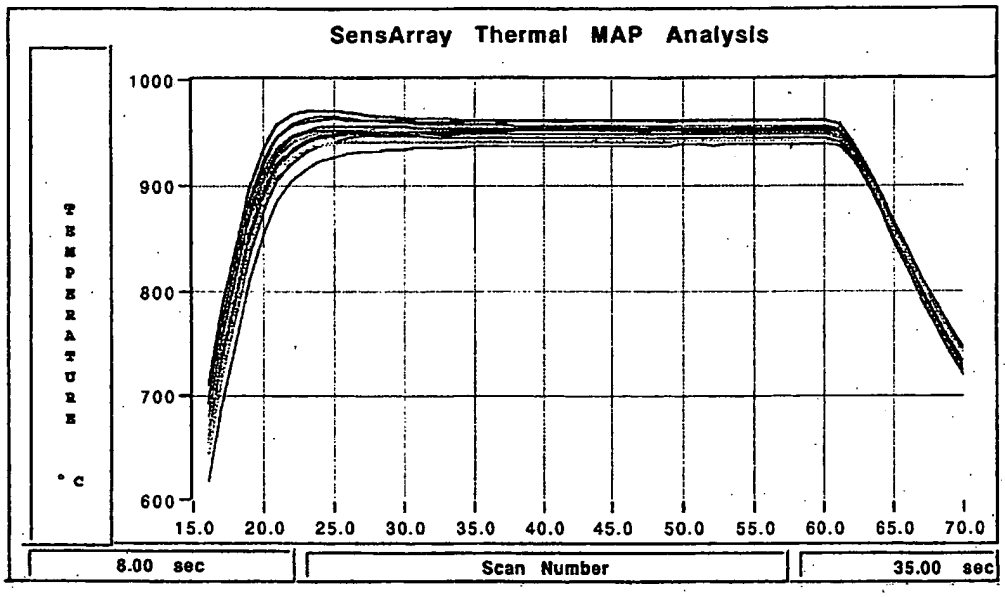


Fig. 7. Temperature profile for "950°C" RTP (ramp rate = 100°C/s) in multi-lamp rapid thermal processor measured with the 17 thermocouple SensArray temperature monitor system.

60°C/s and then to 950°C at 20°C/s. In this case, the maximum wafer temperature was 989°C, an overshoot of 39°C. This was a reduction in the temperature overshoot and across wafer temperature range of more than 40%. The temperature response profile for the two-step ramp process is shown in Fig. 6. The 60°C/s ramp rate process was selected based on the fact that this process eliminated the excessive overlay errors (>0.3 μm) at interconnect metallization, and only a minimal impact of 4% to wafer throughput was incurred. This result differed from that of Feil *et al.* who claimed that a temperature ramp rate reduction did not affect the RTP induced pattern misalignment [12].

A multi-lamp RTP system was also characterized for temperature response using the SensArray system. Fig. 7 shows a temperature profile of the "950°C" RTP in the multi-lamp system for a ramp rate of 100°C/s with the same SensArray

TABLE IV
SENSARRAY CHARACTERIZATION DATA

System	Ramp Rate (C/Second)	Temp. Overshoot (C)	Temp. Range (C)
Single Lamp	100	>70	>120
	60	50	95
	two-step ramp	40	76
Multi Lamp	100	20	49
	60	14	30

thermocouple wafer. For this ramp rate, the temperature overshot the set temperature (950°C) for some wafer perimeter locations. However, the amount of temperature overshoot was drastically lower than that for the single lamp processor. In this case the maximum wafer temperature during the initial transient period was approximately 970°C, an overshoot of 20°C. It was previously shown that for the same ramp rate, the overshoot was greater than 70°C for the single lamp rapid thermal processor. Reducing the ramp rate to 60°C/s in the multi-lamp system reduced the temperature overshoot to approximately 14°C. The overshoot in the multi-lamp processor could be reduced even more by manipulating zone lamp power [13], which was not possible in the single lamp system. A summary of the SensArray characterization of the two rapid thermal processors is shown in Table IV. This table shows the temperature overshoot as well as maximum across wafer temperature range for various temperature ramp rates in both the single and multi-lamp rapid thermal processors (set temperature equal to 950°C). Metal-to-contact misalignment occurred for the case of the 100°C/s ramp rate in the single lamp processor, but not for the reduced ramp rate process, or the 100°C/s ramp rate process in the multi-lamp processor. Therefore, a solution to RTP induced metal-to-contact overlay errors was found, for which both RTP systems could be utilized. A reduced ramp rate process for the single lamp RTP system for temperatures greater than or equal to 950°C ameliorated the misalignment errors, while having only a slight (4%) impact on wafer throughput. The multi-lamp processor was able to accommodate the higher ramp rate of 100°C/s without inducing overlay errors, and was, therefore, capable of maintaining a higher wafer throughput. Therefore, in a 24 hour period, twelve 48 wafer lots could be processed per multi-lamp machine at the 100°C/s ramp rate, while 11.5 wafer lots could be processed per single-lamp machine at the 60°C/s ramp rate. However, this slight decrease of 4% in product throughput for the single-lamp processor was more than compensated for by the approximately 20% increase in product yield.

IV. SUMMARY AND CONCLUSION

Although the state of rapid thermal processing and the equipment have advanced significantly over the years, care must be used in integration of RTP into submicron semiconductor process technologies. This study demonstrated how temperature overshoot in a single lamp rapid thermal processor used for implant activation caused severe wafer distortion and subsequent interconnect metallization-to-contact overlay errors in a 0.85 μm Flash EPROM process based on global alignment photolithography. These RTP induced overlay errors resulted in a functional circuit loss of approximately 20%. The severe wafer distortion and overlay errors occurred for RTP temperatures greater than or equal to 950°C, while no such problems were encountered at temperatures less than or equal to 900°C. However, device design and thermal budget requirements did not allow a simple temperature reduction at this RTP step. The solution to this problem was to either reduce the ramp rate to below 100°C/s for processing in a single lamp system, or to use a multi-lamp processor. In the case of the first solution, a 4% increase in cycle time to process a 48 wafer production lot was incurred. With the use of the multi-lamp system design, no reduction in throughput was incurred, since the ramp rate of 100°C/s could still be used without causing pattern overlay errors.

ACKNOWLEDGMENT

The authors would like to thank Ed Labelle of AMD for collecting the necessary SensArray thermal measurement data in support of this work.

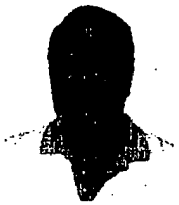
REFERENCES

[1] K. C. Saraswat *et al.*, "Rapid thermal multiprocessing for a programmable factory for adaptable manufacturing of IC's," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, no. 2, pp. 159–175, May 1994.
[2] C. Schaper, "Control of MMST RTP: Repeatability, uniformity, and integration for flexible manufacturing," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, no. 2, pp. 202–219, May 1994.
[3] S. Mehta and D. Hodul, "Process and equipment issues in rapid thermal oxidation (RTO)," in *Proc. Materials Research Soc.—Rapid Thermal Processing of Electronic Materials*, 1987, vol. 92, pp. 95–101.
[4] M. M. Moslehi, K. C. Saraswat, and S. C. Shatas, "Microwave plasma LPCVD of tungsten in a cold-wall lamp heated rapid thermal processor," in *Proc. Materials Research Soc.—Rapid Thermal Processing of Electronic Materials*, 1987, vol. 92, pp. 295–304.
[5] M. C. Oxturk, *et al.*, "Low-pressure chemical vapor deposition of polycrystalline silicon and silicon dioxide by rapid thermal processing," in *Proc. Materials Research Soc.—Rapid Thermal Annealing/Chemical Vapor Deposition and Integrated Processing*, 1989, vol. 146, pp. 109–114.
[6] E. Ma, M. Natan, B. S. Lim, and M.-A. Nicolet, "Comparisons of silicide formation by rapid thermal annealing and conventional furnace annealing," in *Proc. Materials Research Soc.—Rapid Thermal Processing of Electronic Materials*, 1987, vol. 92, pp. 205–212.
[7] S. S. Lee *et al.*, "Optimization of a TiN/$TiSi_2$ p^+ diffusion barrier process," in *Proc. Materials Research Society—Rapid Thermal Annealing/Chemical Vapor Deposition and Integrated Processing*, 1989, vol. 146, pp. 217–222.
[8] H. Ryssel and I. Ruge, *Ion Implantation*. New York: Wiley, 1986, p. 52.
[9] M. M. Moslehi, A. Paranjpe, L. A. Velo, and J. Kuehne, "RTP: Key to future semiconductor fabrication," *Solid-State Technol.*, vol. 37, no. 5, pp. 37–45, May 1994.
[10] P. Singer, "Rapid thermal processing: A progress report," *Semicond. Int.*, p. 64, May 1993.
[11] A. E. Stephens, "Plastic deformation of the silicon wafer," in *Proc. 1st Int. Rapid Thermal Processing Conf.*, Sept. 1993, pp. 94–101.
[12] W. Feil, M. Drew, and J. Moench, "Patterned-induced pattern misregistration after BPSG RTA reflow," in *Proc. 1st Int. Rapid Thermal Processing Conf.*, Sept. 1993, pp. 114–116.
[13] W. Renken, "Process diagnostics using wafer temperature mapping," in *Proc. 1st Int. Rapid Thermal Processing Conf.*, Sept. 1993, pp. 262–266.

James F. Buller (M'84) was born in Mt. Clemens, MI. He received the B.A. degree in mathematics and physics from Albion College, Albion, MI, in 1982, and the M.E. degree in engineering physics from the University of Virginia, Charlottesville, in 1984.

In 1984, he joined Harris Semiconductor, Melbourne, FL, as a Device Engineer. Mr. Buller contributed to the development of SIMOX and bonded wafer material and device technologies, a radiation hardened 256 k CMOS EEPROM process, and a 0.8 μm 256 k SRAM on SOI. In 1992, he joined Advanced Micro Devices, Autin, TX, as a Senior Process Engineer in the Process Development Organization. He has been involved in process development for 1.0 and 0.85 μm UV and Flash EPROM's. He is presently working on the process and device integration of a 0.7 μm CMOS EEPROM PLD technology. Mr. Buller has authored, or co-authored, 13 publications in the areas of semiconductor processing and radiation hardening.

He is a member of the IEEE Electron Devices Society.



M. M. Farahani received the Ph.D. degree in materials science and engineering from the University of Houston, Houston, TX, in 1980.

He joined United Technologies MOSTEK Division in 1980, where he worked back-end as well as front-end process development in the Advanced Research and Development Department. In 1987, he joined Advanced Micro Devices, Austin, TX, where he is presently working as a Senior MTS in the Process Development/Integration Department.

Shyam Garg was born in Indore, India. He received the B.Sc. from the University of Indore in 1969, and the B.E. (electronics) with honors from Jiyajee University in 1973. He later received the M.S. and Ph.D. degrees (solid state electronics), from the University of Cincinnati, Cincinnati, OH, in 1975 and 1981, respectively.

He joined the Intel Corporation, Aloha, OR, in 1981, and worked toward producing the world's first CMOS DRAM. In 1984, he took upon the responsibility of CMOS EPROM development with Texas Instruments, Lubbock, TX. Since 1987, he has been with Advanced Micro Devices, Austin, TX, responsible for the technology development of advanced Flash and and Programmable Logic Devices.

The Effect of Patterns on Thermal Stress During Rapid Thermal Processing of Silicon Wafers

Jeffrey P. Hebb and Klavs F. Jensen

Abstract— **The presence of patterns can lead to temperature nonuniformity and undesirable levels of thermal stress in silicon wafers during rapid thermal processing (RTP). Plastic deformation of the wafer can lead to production problems such as photolithography overlay errors and degraded device performance. In this work, the transient temperature fields in patterned wafers are simulated using a detailed finite-element-based reactor transport model coupled with a thin film optics model for predicting the effect of patterns on the wafer radiative properties. The temperature distributions are then used to predict the stress fields in the wafer and the onset of plastic deformation. Results show that pattern-induced temperature nonuniformity can cause plastic deformation during RTP, and that the problem is exacerbated by single-side heating, increased processing temperature, and increased ramp rate. Pattern effects can be mitigated by stepping the die pattern out to the edge of the wafer or by altering the thin film stack on the wafer periphery to make the radiative properties across the wafer more uniform.**

Index Terms—**Finite element methods, pattern effects, process model, thermal stress, thin film optics, thin film stress.**

I. Introduction

ACHIEVING acceptable across-wafer temperature uniformity during rapid thermal processing (RTP) of silicon wafers has challenged the semiconductor industry for over a decade. Across-wafer temperature gradients lead to thermal stress, which can cause plastic deformation of the wafer if stress levels exceed a critical limit. The plastic deformation of the wafer can give rise to pattern misregistration, where the change in wafer dimensions from one photolithography step to the next causes pattern misalignment and device failure [1], [2]. Wafer warpage can also lead to high levels of stress and generation of defects in dielectric films, degrading device performance [3], [4]. As device dimensions shrink and wafer diameter increases, these problems will become even more critical. It is important to understand the origins of wafer temperature nonuniformity and its effect on thermal stress, so processing problems can be anticipated and solved.

Even if an RTP system delivers acceptable temperature uniformity and thermal stress levels for an unpatterned monitor wafer, the presence of patterns can lead to unacceptable temperature nonuniformity and plastic deformation for the product wafer. This pattern-induced temperature nonuniformity arises from the disparity in radiative properties between the unpatterned wafer periphery and the area where the devices are patterned (see Fig. 1). It has been shown through reactor scale transport modeling [5], [6] and experimentally [7], [8] that this phenomenon can lead to unacceptable temperature nonuniformity. This problem is generally more severe for reactors which do not have dynamic multipoint temperature control, but recent observations have suggested that pattern effects are important even for more sophisticated reactors which have this feature [8].

Direct experimental evidence of plastic deformation during RTP has been demonstrated for unpatterned wafers by Moslehi [9], who showed that the ramp rate and processing temperature were important factors for temperature nonuniformity and plastic deformation for 100-mm diameter wafers. Benetini *et al.* [10] showed that plastic deformation could even occur during rapid thermal processing of 50-mm diameter unpatterned silicon wafers. These works demonstrate the historical difficulty of achieving acceptable temperature uniformity during RTP, even for unpatterned wafers. Vandenabeele *et al.* [8] showed direct experimental evidence that pattern-induced temperature nonuniformity could cause plastic deformation for 150-mm silicon wafers with simple field oxide patterns. Indirect evidence for pattern-induced plastic deformation for product wafers has also been reported in the literature [1], [2]. Feil *et al.* [1] reported that the degree of pattern misregistration after an RTP reflow step was strongly pattern dependent. Buller *et al.* [2] reported similar observations of pattern-dependent misregistration after an RTP contact anneal step.

There have been several efforts to model thermal stress during RTP. Lord [11] modeled the wafer temperature and stress distributions for unpatterned silicon wafers during RTP, using a simple two-dimensional (2-D) reactor scale model and assuming the temperature profiles to be axisymmetric. Erofeev *et al.* [12] modeled the three-dimensional (3-D) temperature and stress distributions of patterned wafers, including the effect of film stress. The temperature distribution was not obtained through a reactor scale transport simulation, but instead a constant radiative heat flux to the wafer was assumed. There has yet to be a study of pattern-induced stress effects in the context of a detailed reactor scale transport model.

The experimental and theoretical work described above has shown that pattern-induced plastic deformation during RTP is a potential problem, but there does not exist a systematic methodology by which pattern-induced stress effects can be evaluated. In this work, we combine a detailed finite-element

Manuscript received March 11, 1997; revised July 20, 1997. This work was supported by the SRC and SEMATECH.

J. P. Hebb is with the Eaton Corporation, Peabody, MA 01960 USA (e-mail: jhebb@bev.etn.com).

K. F. Jensen is with the Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: kfjensen@mit.edu).

Publisher Item Identifier S 0894-6507(98)00335-2.

based reactor transport model, a thin film optics model for predicting radiative properties of patterned wafers, and a simple model for thermal stress to predict the effect of patterns on wafer temperature and stress fields. Using these integrated modeling tools, the effects of wafer patterns, process parameters, and reactor geometry are explored, and potential strategies for minimizing pattern effects are evaluated.

II. Modeling Approach

A. Thermal Modeling

The 2-D axisymmetric reactor scale transport model used in this work has been described by Merchant *et al.* [13]. Starting with a CAD file of the particular RTP system of interest, a finite element mesh is created using a commercial mesh generation package. Radiation exchange factors are calculated by a deterministic ray tracing approach based on the finite element mesh. This approach accounts for multiple diffuse and specular reflections, although in this work it is assumed, for simplicity, that all surfaces are diffuse. Wavelength and temperature dependent properties are taken into account using a three-band approach, described below. The radiative exchange factors are then incorporated into transient finite element fluid flow and heat transfer models to yield temperature and velocity fields, including the time dependent wafer temperature distribution. Thermal cycles are simulated by using a proportional-integral controller to control the wafer center temperature, with the ramp rate and "steady state" processing temperature as inputs.

On the front side of the wafer, there are generally three areas for which it is necessary to calculate the radiative properties (see Fig. 1): the wafer periphery, which is typically covered with homogeneous, optically smooth thin films; the die area, which is optically rough, and consists of a collage of microscale thin film stacks; and the between-die area, which we will assume is covered by homogeneous, optically smooth thin films. The radiative properties for a particular direction and wavelength of each of these areas are calculated using thin film optics, implemented in the form of the matrix method of multilayers [14]. This method accounts for interference effects in thin film stacks, assuming that the layers and interfaces are optically smooth and parallel, and that the lateral dimensions of the sample area are much larger than the wavelength. The wafer periphery area generally satisfies these conditions, and it has been shown that these properties can be predicted accurately at processing temperatures [5], [15]. The die area, on the other hand, generally violates several of the assumptions of thin film optics. Given the pattern density, the radiative properties of a die area are calculated with multilayer theory by assuming that it is an area-weighted mixture, and calculating the radiative properties of each component in the mixture using thin film optics [5]. Even though the use of thin film optics is not strictly valid, there has been experimental evidence suggesting that this simple model provides reasonable approximations for the total radiative properties of the die area [16]. The optical constants of silicon dioxide and silicon nitride are obtained from the literature [17]. The temperature and dopant concentration-dependent optical

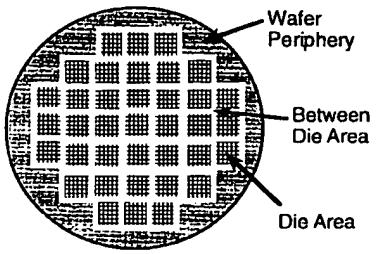


Fig. 1. Schematic showing the three areas of the front-side of a patterned silicon wafer.

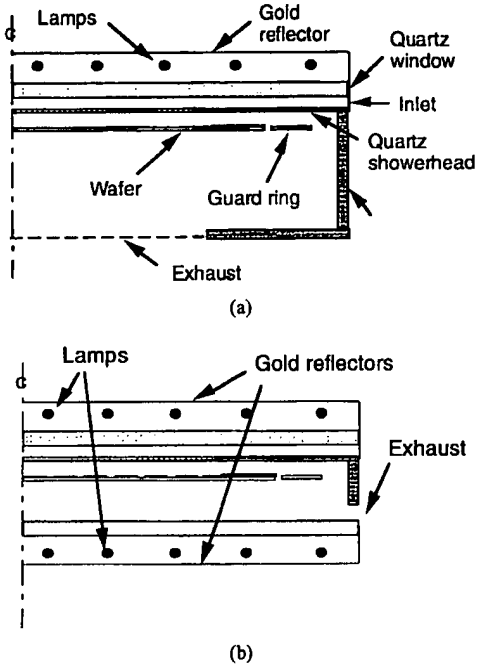


Fig. 2. Generic axisymmetric reactors used as testbeds for pattern effects: (a) Reactor 1, a single-side illumination system; (b) Reactor 2, double-side illumination system. The wafer radius is 100 mm, and the reactors are drawn to scale.

constants of silicon are obtained by a dielectric function model combining correlations from the literature with the Drude model for free carrier absorption [16]. The spectral properties are integrated with respect to direction and wavelength to obtain the total hemispherical radiative properties for each of the three "bands," where each band is specified by a spectral range and a blackbody source temperature [5].

Two axisymmetric reactors are used to explore pattern effects, shown in Fig. 2. Reactor 1 [Fig. 2(a)] is a single-side illumination reactor with five toroidal tungsten halogen lamps and a gold lamphouse reflector on one side, and a large exhaust hole on the other side. Unless otherwise stated, the lamp power is delivered to the front-side of the wafer. Reactor 2 [Fig. 2(b)] is a double-side illumination reactor that has 5 lamps and a gold lamphouse reflector on each side. For the results shown here, the power delivered from the top and bottom lamp banks is equal. The 3-D patterns on the 200-mm diameter wafer (see Fig. 1) are represented by concentric rings of the appropriate length scale. There are seven die areas, each 10-mm wide, six between-die areas, which are 2-mm wide, and the wafer

periphery, which occupies the outer 20 mm of the wafer radius. The radiative properties in each of the three bands can be defined for each area on the wafer. For both reactors, the wafer radius is 100 mm, and the figures are drawn to scale.

Lamp power ratios are optimized for a bare silicon monitor wafer, and these lamp power ratios are then used for the patterned wafer. There is no dynamic adjustment of the lamp power ratios throughout the cycle. This is representative of a typical process for an RTP system having no multipoint dynamic control in which a uniform sheet resistance or silicon dioxide thickness of a blank monitor wafer is used to imply temperature uniformity. For all simulations, the wafer is rotated at 20 rpm, the gas flowrate is 5 slm (N_2), and the operating pressure is 0.01 atmospheres.

B. Stress Modeling

1) Effect of Temperature Nonuniformity: Mechanical stress in the wafer is caused by several factors. Temperature gradients within the wafer lead to thermal stress [18]. Stress is also generated in the wafer due to the presence of thin films, including both intrinsic film stress and stress arising from the mismatch in coefficients of thermal expansion between film and wafer [19]. Finally, gravity can cause mechanical stress, depending on how the wafer is supported [20]. In this analysis, we consider only the effect of temperature nonuniformity, although below there is some discussion on the effect of thin film stress. Regarding the effect of gravity, Huff and Goodall [20] have shown that this is probably not a critical factor for 200-mm diameter wafers, but that it could be a major contributor to mechanical stress if wafer diameter is scaled up to 300 mm. In addition to the above effects, there can be stress concentration at film edges on the die area [21] not accounted for in this work.

The wafer is modeled by assuming that it is an isotropic, elastic, thin, flat circular plate free of external forces which has a uniform temperature through the thickness, and a 2-D axisymmetric radial temperature profile. Our simulations have shown that the maximum temperature difference during the ramp-up across a 750-μm thick wafer is on the order of 1 K, which is not large enough to cause significant levels of thermal stress. Film stress generally causes wafer curvature, which invalidates the assumption of wafer planarity and introduces bending stresses [22] which are not accounted for in the model. The corners at the edges of the die (see Fig. 1) may give rise to stress concentration which is not accounted for in the 2-D analysis. Under these assumptions, the principle stresses in the radial (σ_r) and tangential (σ_θ) directions are given by simple integral expressions [18], which have been used extensively in the literature (e.g., [10], [11])

$$\sigma_r(r,t) = \alpha_s E_s \left\{ \frac{1}{R^2} \int_0^R T(r,t) r\, dr - \frac{1}{r^2} \int_0^r T(r,t) r\, dr \right\} \tag{1}$$

$$\sigma_\theta(r,t) = \alpha_s E_s \left\{ \frac{1}{R^2} \int_0^R T(r,t) r\, dr + \frac{1}{r^2} \int_0^r T(r,t) r\, dr - T(r,t) \right\} \tag{2}$$

where σ_r and σ_θ are the radial and tangential components of stress, r is the radial position, t is time, T is temperature, R is the radius of the wafer, $\alpha_s = 4 \times 10^{-6}$ K^{-1} is the coefficient of thermal expansion of silicon, and $E_s = 1.3 \times 10^5$ MPa is the Young's modulus of silicon [20]. Given the principle stress components, the maximum shear stress for a given radial position and time is calculated using Mohr's circle [23]

$$\tau_m(r,t) = \frac{1}{2} |\sigma_r(r,t) - \sigma_\theta(r,t)|. \tag{3}$$

At high temperatures, silicon behaves like a viscous material, with the yield stress dependent on both temperature and strain rate [18]. The yield stress is also strongly dependent on initial dislocation density, dissolved oxygen concentration, and precipitated oxygen concentration. For example, both a precipitated oxygen concentration of 4×10^{17} cm^{-3} can lower the yield stress by more than a factor of two, and an initial dislocation density of 5×10^{-4} cm^{-3} can decrease the yield stress by a factor of two [20]. These parameters depend on the manufacturing of the wafer, as well as the process history and mechanical handling of the wafer [20]. These effects are complex and difficult to predict, and so we assume that the silicon is dislocation and oxygen free, which will generally overpredict the yield stress of the silicon. We use the correlation of Widmer and Rehwald [24] to describe the yield stress in shear of dislocation, oxygen-free silicon, which fits the data of Patel and Chaudhuri [25]

$$\tau_y(r) = A \left(\frac{\dot{\epsilon}(r)}{\dot{\epsilon}_0} \right)^{1/n} \exp\left(\frac{\Delta E}{kT(r)} \right) \tag{4}$$

where τ_y is the yield stress, $A = 1815$ Pa, $\dot{\epsilon}$ is the axial strain rate, $\dot{\epsilon}_0 = 1 \times 10^{-3}$ s^{-1}, $n = 2.45$, $\Delta E = 1.07$ eV is the activation energy for dislocation movement [26], k is Boltzmann's constant, and T is the temperature in degrees Kelvin. The axial strain rate is calculated from the shear strain rate, which is readily extracted from the transient simulations. Based on a scaling analysis of creep rates at steady state and the use of (4) by other workers to fit experimental data on wafer warpage [10], [24], a minimum axial strain rate of 4×10^{-6} s^{-1} is used in this work. The yield stress will decrease exponentially with temperature, and decrease with the square root of strain rate, both of which have important implications for plastic deformation during RTP. The yield stress is shown in Fig. 3, which shows that temperature is the first-order effect, with strain rate playing a significant but secondary role.

Using the maximum shear stress criterion for plastic deformation [23], it is assumed that there will be local plastic deformation when

$$\Omega(r) = \frac{\tau_m(r)}{\tau_y(r)} > 1 \tag{5}$$

where Ω is referred to as the stress ratio. It should be noted that since (1) and (2) assume elastic behavior, the analysis is invalid for stress ratios greater than 1. However, the magnitude of the stress ratio can still be used to give a qualitative approximation of the degree of plastic deformation. Given the simplifying assumptions that are made for calculation of shear stress, and the uncertainties in the yield stress, the calculations below

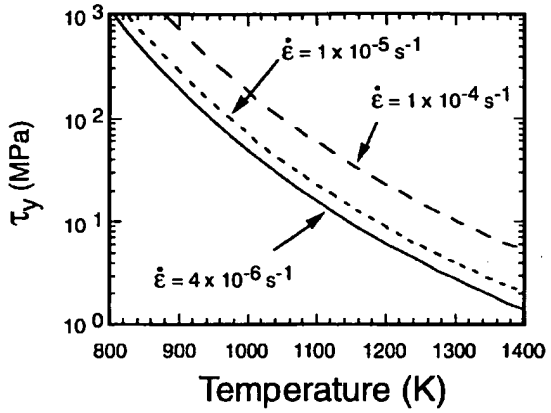


Fig. 3. The temperature dependence of the yield stress in shear of dislocation, oxygen-free silicon for various strain rates.

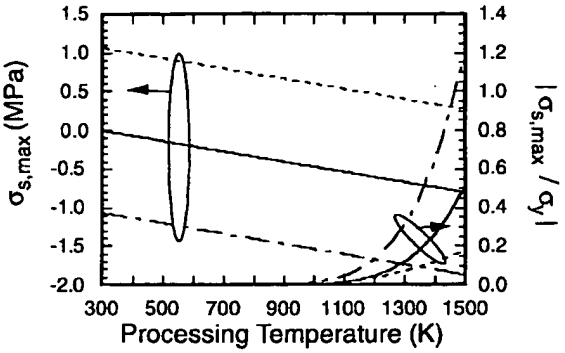


Fig. 4. The maximum normal stress in a 750-μm thick silicon wafer and the ratio of maximum stress to yield stress caused by a 0.2-μm Si_3N_4 film, for various values of intrinsic film stress: +1 GPa (solid line), 0 (dotted line), and −1 GPa (dot-dash line).

should be seen as a guide to whether or not there will be plastic deformation.

2) Effect of Thin Film Stress: Although we are ignoring the effect of film stress, it is useful to assess the validity of this negation. At room temperature, there is typically some intrinsic stress in the film, where both the sign and magnitude of the intrinsic stress depend heavily on the processing conditions of the film [19]. Stress also arises as the wafer is heated, due to the mismatch in coefficients of thermal expansion between film and wafer. Here we assume that there is a single blanket film on one side of the wafer, with the other side bare. Ignoring any effects at the edge of the wafer, the maximum normal stress within the wafer is given by [27]

$$\sigma_{s,\max} = -4\frac{t_f}{t_s}\left[\sigma_{f,\text{int}} + \frac{E_f}{1-\nu_f}(\alpha_s - \alpha_f)(T - T_o)\right] \quad (6)$$

where $\sigma_{s,\max}$ is the stress in x and y directions parallel to the film–wafer interface; t_f and t_s are the film and wafer thicknesses, respectively; $\sigma_{f,\text{int}}$ is the intrinsic stress in the film; E_f is the Young's modulus of the film; ν_f is the Poisson's ratio of the film; α_s and α_f are the coefficients of thermal expansion of the wafer and film, respectively; and T_o is the reference temperature. The negative sign indicates that a compressive stress in the film will result in a tensile stress in the substrate. The factor of 4 arises from the bending stresses that the film causes. To assess the effects for plastic deformation, the ratio of the maximum normal stress in the wafer is compared to the normal yield stress of silicon, σ_y, which is readily extracted from (4).

Fig. 4 shows the maximum stress in the wafer, and the normal stress ratio as a function of temperature for a silicon nitride film of 0.2 μm. Depending on how the film was deposited, the intrinsic stress can range from approximately −1 GPa to +1 GPa [28], [29]. We plot the wafer stress for intrinsic stresses of +1 GPa, −1 GPa, and 0. For the nitride film we assume $\alpha_f = 2.5\times10^{-6}\ \text{K}^{-1}$ [19], $E_f = 3\times10^5$ MPa [29], $\nu_f = 0.28$, and we assume a wafer thickness of 750 μm. Because the thermal expansion coefficient of silicon nitride is smaller than that of the silicon, increasing the temperature tends to give a compressive stress in the wafer. Therefore, an increase in temperature decreases the stress in the wafer if the film has a tensile intrinsic stress, and increases it if the film has a compressive intrinsic stress. For all three cases, the stress ratio increases with temperature due to the temperature dependence of the yield stress of silicon. The stress ratio only exceeds unity for the wafer with film that is intrinsically tensile, but even then it is only at very high temperatures. Calculations done for a 0.5 μm thick silicon dioxide film show similar trends and magnitudes for the ratio of maximum stress to yield stress [30]. These calculations show that at normal RTP temperatures, thin film stress could play a significant role in plastic deformation, although this is probably not a first-order effect.

There are many factors that this analysis does not take into account. For real wafers, there are typically blanket films on the wafer periphery, and microscale thin film patches in the die area, which would tend to compensate for the blanket films on the back-side. Furthermore, a complete analysis would consider the total mechanical stress in the wafer, which is a result of superimposing stress from thin films, temperature gradients, and gravity.

III. Results and Discussion

The thermal cycle and patterns used for the simulations are chosen to represent the rapid thermal annealing of a silicon wafer for shallow junction formation. For all simulations, the wafer center temperature is increased at a defined ramp rate, held at the "steady state" processing temperature for 30 s, and then the lamps are switched off. Unless otherwise stated, the results shown are for a ramp rate of 80 K/s and a processing temperature of 1273 K. Table I gives a description of three wafer patterns to be investigated and the radiative properties calculated for a wafer temperature of 1273 K in each of the three bands. The first two bands, denoted by subscripts 1 and 2, indicate the emittance of the wafer for wavelengths below and above 4.0 μm, respectively. The third band, denoted by subscript ℓ, indicates the absorptance of the wafer for energy incident from the lamps, which are assumed to be at a temperature of 3000 K [4]. The wafer is undoped and 750 μm thick. For all patterns, the die area has three component film stacks (top film first): 1) 20% bare silicon, 2) 30% 0.3 μm polysilicon/0.5 μm SiO_2, and 3) 50% 0.5 μm

TABLE I
RADIATIVE PROPERTIES OF PATTERNED SIDE OF THE WAFER FOR THREE ANNEALING PATTERNS. KEY FOR THE LAYERS: POLY = 0.3-μm POLYSILICON, NITRIDE = 0.2-μm Si_3N_4, FIELD OXIDE= 0.5 μm SiO_2, GATE OXIDE = 100-Å SiO_2

PATTERN	AREA	LAYERS	ε_1	ε_2	α_ℓ
all patterns	die	20 % bare Si 30 % poly / field oxide 50 % field oxide	0.69	0.67	0.67
	between die	field oxide	0.82	0.71	0.75
	back-side	bare silicon	0.66	0.68	0.65
pattern 1	border	nitride / poly / gate oxide	0.81	0.71	0.80
pattern 2	border	same as die area	0.69	0.67	0.67
pattern 3	border	poly / gate oxide	0.68	0.65	0.65

SiO_2. For all patterns, the between-die area is covered with 0.5 μm SiO_2. For pattern 1, the wafer periphery has a thin film stack consisting of (top film first): 0.2 μm Si_3N_4/0.3 μm polysilicon/100 Å SiO_2. For pattern 2, the wafer periphery has the die pattern stepped out to the wafer edge. For pattern 3, the wafer periphery is the same as pattern 1 except that the Si_3N_4 layer is removed. We assume that the back-side is bare silicon for all patterns.

Fig. 5 shows results for the single-side illumination reactor for a bare monitor wafer and a wafer with pattern 1. Fig. 5(a) shows the wafer center temperature and the absolute value of the difference between the maximum and minimum temperature across the wafer throughout the 60 s cycle. For the bare wafer, the maximum temperature difference is approximately 40 K, occurring during the ramp, and settles down to a maximum temperature difference less than 5 K during the steady state. Because of wafer emission to the exhaust hole, more power has to be delivered to the center of the wafer to achieve uniform temperature at steady state. During the ramp, the wafer is not hot enough for the emission to the exhaust hole to compensate for the extra power delivered to the center, and so the center gets hotter than the edge. For the patterned wafer, the increased absorptance of the wafer periphery slightly offsets this center overheating during the ramp, but causes an unacceptable 25 K edge overheating during the steady state part of the cycle. For the next generation of integrated circuit (IC) chips, it is estimated that a temperature uniformity of 5 K or less during RTP will be required [31].

Fig. 5(b) shows the wafer center temperature and the maximum stress ratio across the wafer throughout the 60 s cycle. For the bare wafer, the model predicts no plastic deformation, whereas for the patterned wafer, the stress ratio is almost 2 throughout most of the cycle. Since the maximum shear stress exceeds the yield stress by almost a factor of two, we would expect to see severe plastic deformation of the patterned wafer. For both the patterned and unpatterned wafer, the peak stress ratio does not occur at the same time as the peak temperature nonuniformity. The peak temperature nonuniformity takes place at a wafer center temperature of about 900 K, when the yield stress of silicon is relatively large (see Fig. 2). The maximum stress ratio occurs near the end of the ramp, when the wafer is near the "steady state"

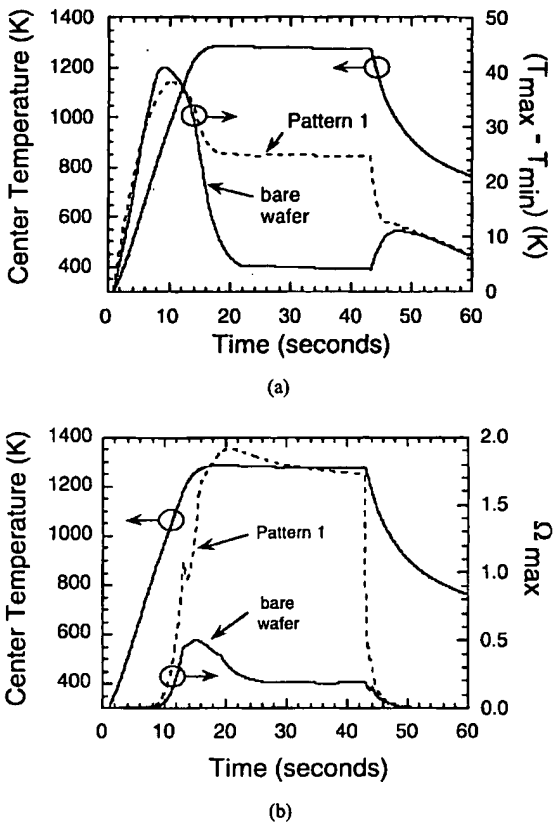


Fig. 5. (a) Wafer center temperature and maximum temperature difference across the wafer during a simulated implant anneal cycle in the single-side illumination reactor, for bare monitor a wafer and a wafer with pattern 1. (b) Wafer center temperature and maximum stress ratio across the wafer for the same thermal cycle. The trajectories of the center temperatures for the bare and patterned wafers are indistinguishable.

temperature, and the yield stress is low due to the exponential dependence on temperature. Also, the strain rate is decreasing at the end of the ramp due to a decrease in ramp rate, which further decreases the yield stress.

Fig. 6 shows the radial temperature and stress ratio profiles at $t = 20$ s, when the stress ratio peaks for the wafer with pattern 1. It shows that the higher absorptance of the wafer periphery is the cause of the edge overheating during the steady state part of the cycle, as discussed above. The temperature gradient caused by the higher absorptance of the wafer periphery is large enough to cause plastic deformation in this area of the wafer. The higher absorptance between die areas leads to local hot spots, but these temperature gradients are not large enough to cause plastic deformation in these areas. One hundred biquadratic finite elements were used to resolve the radial temperature variation. The lack of effect of local hot spots due to scribe lines is an encouraging result, since it suggests that the most detrimental temperature nonuniformities take place on length scales that are controllable through lamp tuning.

Fig. 7 shows analogous results for the double-sided illumination system. Fig. 7(a) shows that during the ramp, the peak maximum temperature difference across the wafer is much higher for Reactor 2 than for Reactor 1. To achieve

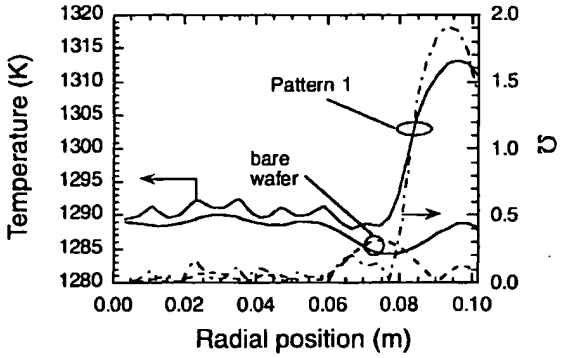


Fig. 6. Spatial variation of temperature and stress ratio at $t = 20$ s, the time of peak stress ratio, for the simulation shown in Fig. 5.

acceptable steady state temperature uniformity in the double-side illumination system, more radiative flux must be provided to the edge of the wafer to compensate for the greater emissive losses near the edge. During the ramp, the radiative losses are much less than during steady state due to the lower edge temperature, and so the increased power to the edge causes it to be much hotter than the center. The higher absorptance of the wafer periphery exacerbates the problem of edge overheating during the ramp, the maximum temperature difference peaking at about 150 K. During the steady state, the temperature uniformity of the patterned wafer is nearly equal to that of the bare wafer. Because less power is being delivered to the patterned side in this system, the disparities in lamp absorptance do not lead to the severe pattern effects seen for the single-side illumination system. Also, the lamps are tuned in such a way that the bare wafer has a slightly cooler edge during steady state, offsetting the effect of edge overheating due to the higher absorptance of the wafer periphery.

Fig. 7(b) shows the wafer center temperature and the maximum stress ratio across the wafer throughout the 60 s cycle. Even the bare wafer experiences plastic deformation near the end of the ramp due to edge overheating. To avoid plastic deformation in this system, adjustment of the lamp power ratios during the cycle would be necessary. However, the peak maximum stress ratio for the patterned wafer is only about 30% larger than for the bare wafer, compared to a factor of 4 for single-side illumination reactor. By the time the wafer gets to temperatures high enough to give a low yield stress, the maximum temperature difference for the patterned wafer is only slightly higher than that for the bare wafer. Throughout the steady state, the maximum stress ratio is less than unity for both the bare and patterned wafers. Fig. 8 shows the radial temperature and stress ratio profiles at the time when the stress ratio peaks. The edge of the patterned wafer gets hotter than that of the bare wafer, but the gradients are not significantly steeper. This further illustrates why the pattern does not increase the peak stress ratio drastically.

To mitigate pattern effects, the process engineer could adjust the lamp power ratios until the observed problem disappears. Problems with temperature nonuniformity are more difficult and costly to assess using the product wafer than using a blank monitor wafer, making the adjustment of lamp power

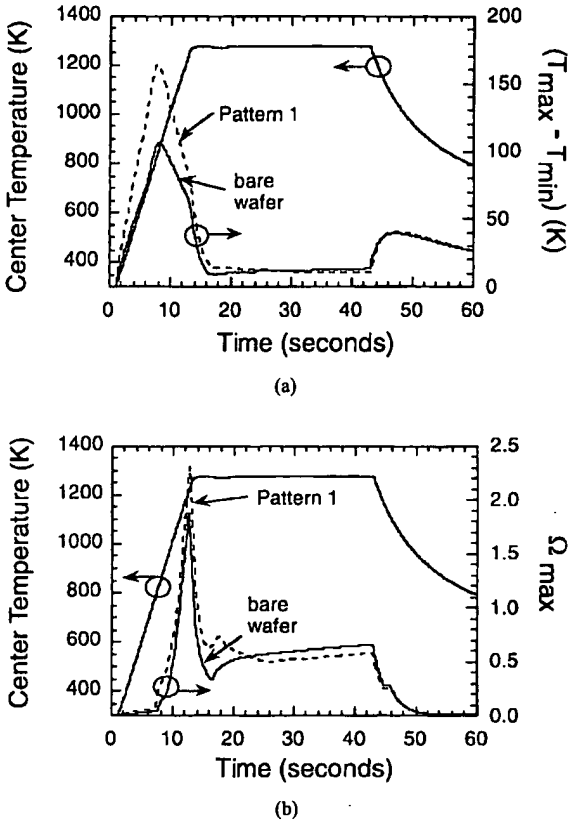


Fig. 7. (a) Wafer center temperature and maximum temperature difference across the wafer during a simulated implant anneal cycle in the double-side illumination reactor, for bare monitor a wafer and a wafer with pattern 1; (b) Wafer center temperature and maximum stress ratio across the wafer for the same thermal cycle. The trajectories of the center temperatures for the bare and patterned wafers are indistinguishable.

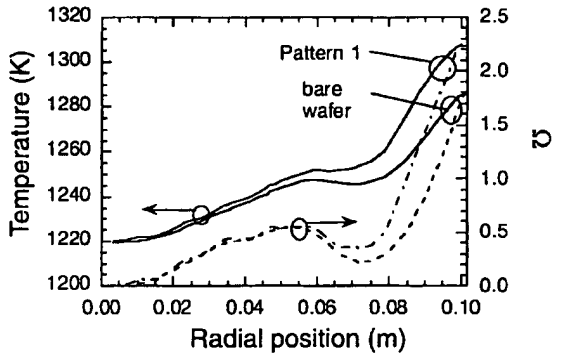
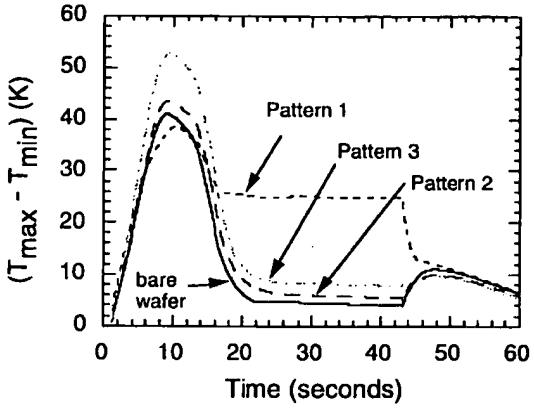
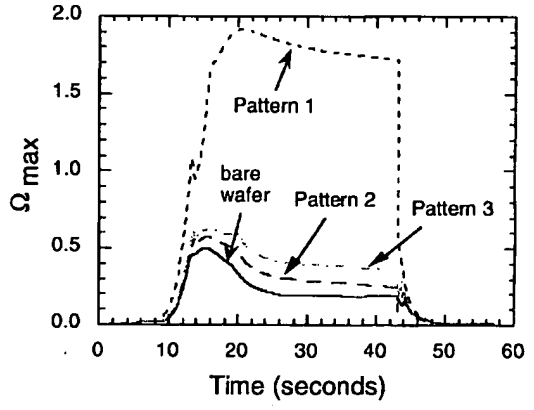


Fig. 8. Spatial variation of temperature and stress ratio at $t = 12$ s, the time of peak stress ratio, for the simulation shown in Fig. 7.

ratios to optimize the product wafer an undesirable approach. An alternative is for the engineer to alter the wafer to try to reduce the disparity between the radiative properties of the die area and wafer periphery. This can be done by either stepping the die pattern out to the edge of the wafer, or by removing or adding layers to the wafer periphery. The effectiveness of these two approaches is shown in Fig. 9, with pattern 2 having the die pattern stepped out to the wafer edge, and



(a)



(b)

Fig. 9. (a) Maximum temperature difference across the wafer, and (b) maximum stress ratio for a bare wafer and wafers with patterns 1, 2, and 3 for a simulated implant anneal in the single-side illumination reactor.

pattern 3 having the nitride layer removed from the wafer periphery. Fig. 9(a) shows that throughout most of the cycle, the temperature nonuniformity for patterns 2 and 3 is much less than those of pattern 1, and even approaches that of the bare wafer. This can be explained by the improved uniformity of the radiative properties across the wafer achieved by both patterns 2 and 3 (see Table I). The maximum transient nonuniformity during the ramp is larger for pattern 3 because the slightly smaller absorptance of the wafer periphery exacerbates the problem of center overheating. Fig. 9(b) shows that both of these strategies are successful in avoiding plastic deformation throughout the entire cycle. The strategy of stepping the die pattern out to the edge is slightly more effective, but this strategy would be accompanied by a significant reduction in throughput. Removing the nitride layer, on the other hand, would add only one extra processing step.

Fig. 10 shows the effect of processing temperature for the single-side illumination reactor. For each processing temperature, the peak value of the maximum stress ratio throughout the cycle is extracted from the simulation. Results are shown for front-side heating of the wafers with patterns 1, 2 and 3, and for back-side heating of the wafer with pattern 1. The

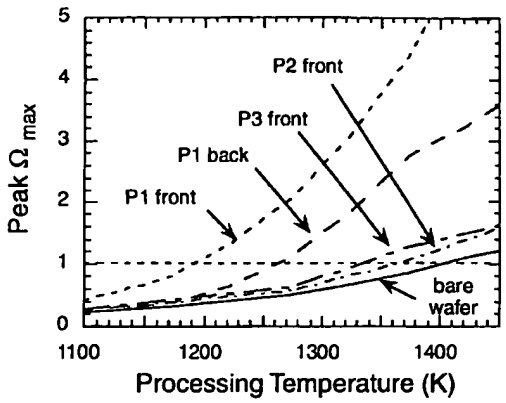


Fig. 10. The effect of processing temperature on the peak maximum stress ratio for the single-side illumination reactor, for the same implant anneal cycle shown in Fig. 5. Results are shown for patterns 1, 2, and 3 for front-side heating, and for pattern 1 for back-side heating. The maximum allowable processing temperature for deformation-free processing is denoted when the peak maximum stress ratio exceeds unity.

general trend is the same for all patterns: plastic deformation becomes more likely as processing temperature is increased due to the temperature dependence of the yield stress. In order to have deformation-free processing for pattern 1 for front-side heating, the processing temperature would have to be decreased to 1200 K. Back-side heating has been explored by several authors [5], [6] as a possible strategy for reducing pattern effects. In this case, back-side heating increases the allowable processing temperature to over 1250 K. However, it was shown by Hebb and Jensen [5] that back-side heating does not always improve temperature uniformity because disparities in emittance can lead to localized cooling. The largest increase in allowable processing temperature comes from altering the radiative properties of the wafer periphery, which raises the allowable processing temperature to over 1300 K.

Fig. 11 shows the effect of ramp rate for the single-side heating system with pattern 1. Although the ramp rate nearly doubles the maximum temperature difference by increasing the ramp rate from 50 to 125 K/s, the peak maximum stress ratio only increases by approximately 10%. For all three ramp rates, the peak maximum stress ratio does not coincide with the peak maximum temperature difference because of the effect of wafer temperature on yield stress. For this reactor with this pattern, the ramp rate is a significant, but secondary effect, compared to the effects of the radiative properties of the wafer periphery and process temperature.

Although no direct comparison of model predictions and data has been done, the trends predicted by the model are generally consistent with experimental observations. Feil *et al.* [1] and Buller *et al.* [2] reported that photolithography overlay errors became larger with increasing processing temperature, consistent with Fig. 10. Buller *et al.* [2] also observed that overlay errors increased with increasing ramp rate, a trend which agrees with our simulations. Finally, the results in Fig. 2(b) for pattern 3 are consistent with the observations of Feil *et al.* [1], who reported that removing one of the layers from the stack on the wafer periphery made a large difference in photolithography overlay errors.

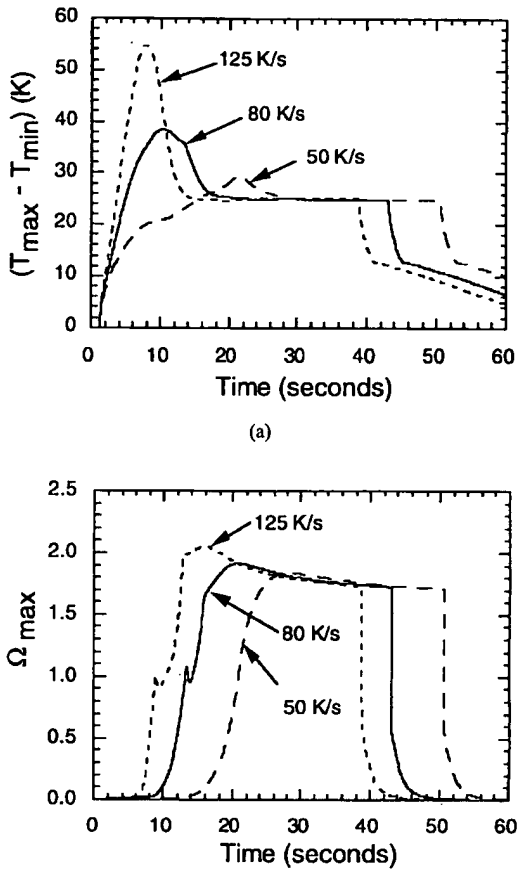


Fig. 11. The effect of ramp rate on (a) maximum temperature difference across the wafer. (b) Maximum stress ratio for a bare wafer and a wafer with pattern 1, for a 30 s implant anneal at 1273 K in the single-side illumination reactor.

IV. Conclusion

This work has shown that pattern-induced temperature gradients can cause plastic deformation during RTP. Of the two reactors studied, pattern effects were more severe for the single-side illumination system, although the double-side illumination system required dynamic control of the lamp power ratios to achieve deformation-free processing, even for an unpatterned wafer. The simulations show that the temperature gradients responsible for plastic deformation take place on a length scale that is controllable by lamp power tuning. Simulations also show that pattern-induced plastic deformation becomes more severe as processing temperature and ramp rate increase. Pattern-induced plastic deformation can be reduced either by stepping the die pattern out to the edge of the wafer, or by altering the stack on the wafer periphery to make the radiative properties across the wafer more uniform. Altering the thin film stack on the wafer periphery is a promising strategy because it would not drastically reduce throughput, but modeling is required to assist the process engineer in deciding which layers to add or subtract. The above results are specific to the parameters used in the simulations. This work provides a methodology by which pattern-induced stress effects can be assessed *a priori* for other cases, and strategies for minimizing these effects can then be explored.

As wafer diameter is scaled up to 300 mm, and the device dimensions shrink further into the sub-micron regime, issues regarding temperature uniformity and thermal stress will become even more critical for RTP. Achieving uniform temperature during RTP will be more difficult, tolerances for misalignment during photolithography will be stricter, and device performance will be more sensitive to stress-induced defects. With this scale up, the effects of thin films and gravity on mechanical stress will become more important, and more sophisticated models will be required.

Acknowledgment

The authors would like to thank SRC and SEMATECH for their support.

References

[1] B. Feil, M. Drew, and J. Moench, "Patterned-induced pattern misregistration after BPSG RTA reflow," in *Proc. 1st Int. Rapid Thermal Processing Conf.*, Scottsdale, AZ, Sept. 8–10, 1993, pp. 114–116.

[2] J. F. Buller, M. Farahani, and S. Garg, "RTA induced overlay errors in a global alignment stepper technology," in *Proc. 2nd Int. Rapid Thermal Processing Conf.*, Monterey, CA, Aug. 31–Sept. 2, 1994, pp. 52–56.

[3] R. P. S. Thakur, N. Chhabra, and A. Ditali, "Effects of wafer bow and warpage on the integrity of thin gate oxides," *Appl. Phys. Lett.*, vol. 64, pp. 3428–3430, 1994.

[4] D. L. Chapek, R. A. Weimer, K. F. Scheugraf, A. Ahmad, R. P. S. Thakur, and R. Singh, "Correlation between thermal stress and the performance of devices processed by RTP," in *Proc. 3rd Int. Rapid Thermal Processing Conf.*, Amsterdam, The Netherlands, Aug. 30–Sept. 1, 1995, p. 281.

[5] J. P. Hebb and K. F. Jensen, "The effect of multilayer patterns on temperature uniformity during rapid thermal processing," *J. Electrochem. Soc.*, vol. 143, pp. 1142–1151, 1996.

[6] P. Vandenabeele and K. Maex, "Temperature nonuniformity during rapid thermal processing of patterned wafers," in *Proc. SPIE*, 1989, vol. 1189, pp. 89–103.

[7] J. Keuhne, S. Hattangady, and M. Pas, "Effects of patterned films on the uniformity of rapid thermal oxidation," in *Proc. 4th Int. Rapid Thermal Processing Conf.*, Boise, ID, Sept. 11–14, 1996, pp. 417–420.

[8] P. Vandenabeele, K. Maex, and R. De Keersmaecker, "Impact of patterned layers on temperature nonuniformity during rapid thermal processing," *Mater. Res. Soc. Proc.*, vol. 146, pp. 149–160, 1989.

[9] M. M. Moslehi, "Process uniformity and slip dislocation patterns in linearly ramped-temperature transient rapid thermal processing of silicon," *IEEE Trans. Semiconduct. Manufact.*, vol. 2, pp. 130–140, 1989.

[10] G. Benetini, L. Correra, and C. Donolato, "Defects introduced in silicon wafers during rapid isothermal annealing: Thermoelastic and thermoplastic effects," *J. Appl. Phys.*, vol. 56, pp. 2922–2929, 1984.

[11] H. A. Lord, "Thermal stress analysis of semiconductor wafers in a rapid thermal processing oven," *IEEE Trans. Semiconduct. Dev.*, vol. 1, pp. 105–114, 1988.

[12] A. F. Erofeev, T. M. Makhviladze, A. V. Panjukhin, O. S. Volchek, and O. Adetutu, "Simulation of thermal warpage and stress in patterned wafers during RTP," in *Proc. 4th Int. Rapid Thermal Processing Conf.*, Boise, Idaho, Sept. 11–13, 1996, pp. 342–346.

[13] T. P. Merchant, J. V. Cole, K. L. Knutson, J. P. Hebb, and K. F. Jensen, "A systematic approach to simulating rapid thermal processing systems," *J. Electrochem. Soc.*, vol. 143, pp. 2035–2043, 1996.

[14] P. Yeh, *Optical Waves in Layered Media*. New York: Wiley, 1988.

[15] P. Timans, "The effect of coatings on the emissivity of silicon," in *Proc. 2nd Int. Rapid Thermal Processing Conf.*, Monterey, CA, Aug. 31–Sept. 2, 1994, pp. 186–193.

[16] J. P. Hebb and K. F. Jensen, "Pattern induced temperature nonuniformity during rapid thermal processing," in *Proc. 4th Int. Rapid Thermal Processing Conf.*, Boise, ID, Sept. 11–13, 1996, pp. 34–39.

[17] E. D. Palik, *Handbook of Optical Constants of Solids*. New York: Academic, New York, 1985.

[18] B. A. Boley and J. H. Weiner, *Theory of Thermal Stresses*. Malabar, FL: Krieger, 1985.

[19] M. Ohring, *The Materials Science of Thin Films*. San Diego, CA: Academic, 1992.
[20] H. R. Huff and R. K. Goodall, "Challenges and opportunities for dislocation free silicon wafer fabrication and thermal processing: An historical review," *Proc. 3rd Int. Rapid Thermal Processing Conf.*, Amsterdam, The Netherlands, Aug. 30–Sept. 1, 1995, pp. 9–40.
[21] S. M. Hu, "Stress related problems in silicon technology," *J. Appl. Phys.*, vol. 70, pp. R53–R80, 1991.
[22] L. D. Dyer, H. R. Huff, and W. W. Boyd, "Plastic deformation in central regions of epitaxial silicon slices," *J. Appl. Phys.*, vol. 42, pp. 5680–5688, 1971.
[23] W. H. Bowes, L. T. Russell, and G. T. Suter, *Mechanics of Engineering Materials*. New York: Wiley, 1984.
[24] A. E. Widmer and W. Rehwald, "Thermoplastic deformation of silicon wafers," *J. Electrochem. Soc.*, vol. 133, pp. 2403–2409, 1986.
[25] J. R. Patel and A. R. Chaudhuri, "Macroscopic plastic properties of dislocation-free germanium and other semiconductor crystals. I. Yield behavior," *J. Appl. Phys.*, vol. 34, pp. 2788–2799, 1963.
[26] W. Schroter, H. G. Brion, and H. Siethoff, "Yield point and dislocation mobility in silicon and germanium," *J. Appl. Phys.*, vol. 54, pp. 1816–1820, 1983.
[27] S. P. Baker and P. H. Townsend, "Mechanical properties of thin films," *Mat. Res. Soc. Fall Meeting Tutorial*, 1995.
[28] P. N. Kember, S. C. Liddel, and P. Blackborrow, "Characterization of plasma deposited silicon nitride as applied to novel MOS structures," *Semicond. Int.*, vol. 8, p. 8, 1985.
[29] E. W. Hearn, D. J. Werner, and D. A. Dooney, "Film induced stress model," *J. Electrochem. Soc.*, vol. 8, p. 1749, 1986.
[30] J. P. Hebb, "Pattern effects in rapid thermal processing," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1997.
[31] *The National Technology Roadmap for Semiconductors*, Semiconductor Industry Assoc., San Jose, CA, 1994.

Jeffrey P. Hebb received the B.Sc. degree in mechanical engineering from the Technical University of Nova Scotia, Halifax, N.S., Canada, and the M.Sc. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge.

He is a Process Engineer with Eaton Corporation, Thermal Processing Systems, Peabody, MA. His interests are the design and understanding of thermal processing systems, including fundamentals of thermal radiation of multilayer structures. He is the author of several publications on RTP simulations.

Dr. Hebb received the best paper award from the Second International Rapid Thermal Processing Conference. He held a Canadian National Science and Post Graduate Study Award from 1992 to 1994.

Klavs F. Jensen received the M.Sc. degree in chemical engineering from the Danish Technical University and the Ph.D. degree from the University of Wisconsin, Madison.

He is the Lammot du Pont Professor of Chemical Engineering and Professor of Materials Science and Engineering at the Massachusetts Institute of Technology, Cambridge. His research interests revolve around chemistry and transport phenomena related to processing of micro- and nano-structured materials for electronic and optical applications, including organometallic chemical vapor deposition (OMCVD). A long-term emphasis has been the development of detailed process models for the design and control of CVD and rapid thermal processing (RTP) equipment used in the microelectronics industry. He is the co-author of more than 240 publications, including several edited volumes.

Dr. Jensen is the recipient of several awards, including the Electrochemical Society Young Authors' Award in Solid State Science and Technology, a National Science Foundation Presidential Young Investigator Award, a Guggenheim Fellowship, and the Allan P. Colburn and Charles C. M. Stine Awards of the American Institute of Chemical Engineers.

A New Approach to Correlating Overlay and Yield

Moshe E Preil * and John F. M. McCormack +

KLA-Tencor Yield Management Consulting
* One Technology Drive, Milpitas, CA 95035
+ Unit 5, Alderstone Business Park, MacMillan Road, Livingston EH54 7DF Scotland

ABSTRACT

Integrated circuit design rules are defined with a given overlay tolerance, but the exact correlation between measured overlay on product wafers and die yield is notoriously difficult to quantify. Interest in better quantifying this relationship is not merely academic. The ability to shrink the overlay design rule by even a few nanometers would allow more good die to be printed on every product wafer, providing a substantial economic benefit. Conversely, if the actual distribution of overlay errors across a wafer is slightly worse than anticipated in the design rules, the resulting shortfall in yield would be difficult to identify and correct.

Traditional linear regression analysis produces poor results for overlay to yield correlation. Product overlay data is usually quite limited, necessitating the use of wafer or even lot level averages which obscures the finer details of overlay variation. Statistical fluctuations in other parameters, most notably critical dimensions, and across wafer yield variations due to other process steps also obscures the correlation. Finally, linear regression is poorly suited to the problem of correlating a continuous predictor (overlay) with a dichotomous (pass/fail) response such as yield.

In this paper, we present a new approach to this problem which addresses many of the shortcomings of traditional linear regression. We employ an overlay simulation model to predict the maximum overlay error for each die, and compare the die level overlay values to yield results. This change from wafer level averages to die level correlation will be shown to be a powerful analysis tool and significantly improves the validity of the correlation. We will demonstrate that die level correlation helps to identify the overlay induced yield losses and removes the masking effects of other random and systematic sources of yield loss.

Keywords: Overlay, yield, die level correlation, process shrinks

1. INTRODUCTION

The acceleration of the SIA Roadmap[1] in recent years is just one more indication of the ever increasing pressure to shrink integrated circuit die sizes while improving their speed and capability. While most of the attention is given to shrinking linewidths, overlay also plays a vital role in determining die size, and hence the number of die which can be fabricated on each wafer. When fabs invest billions of dollars in improved capacity to print smaller geometries, they must at the same time improve their overlay capability in order to maximize the number of die they produce. Failure to fully utilize expensive wafer area can be economically ruinous in an era of hyper competitive technology and collapsing prices for most devices.

Overlay errors can harm profitability in two entirely opposite ways. If the tolerances are not tight enough, the die can not be shrunk, and the number of die per wafer can not be increased. If the tolerances are too tight, yield losses will escalate. In either case, the fab will not be operating efficiently. As a result, overlay optimization is truly a delicate balancing act. In setting overlay design rules, care must be taken not to set the overlay tolerance tighter than the actual fab capability under realistic manufacturing conditions. The yield fall off due to overlay issues is quite dramatic when the design rule approaches the capability limit. Fig. 1 shows simulated results for net good die per wafer vs. overlay tolerance for three different levels of overlay capability; the data is derived from the published results by Arnold and Greeneich[2]. Clearly, one must avoid setting the tolerance to the left of the yield peak to avoid a drastic reduction in yield. On the other hand, setting the tolerance too loose will result in a high percentage yield, but far fewer die per wafer. Relaxing the specifications to make manufacturing easier is not an option given the financial realities of the IC industry. The key is to optimize fab performance and match the design rules as closely as possible to the resulting capability in order to produce the maximum number of good die per wafer.

Part of the SPIE Conference on Metrology, Inspection, and Process Control for Microlithography XIII • Santa Clara, California • March 1999
SPIE Vol. 3677 • 0277-786X/99/$10.00

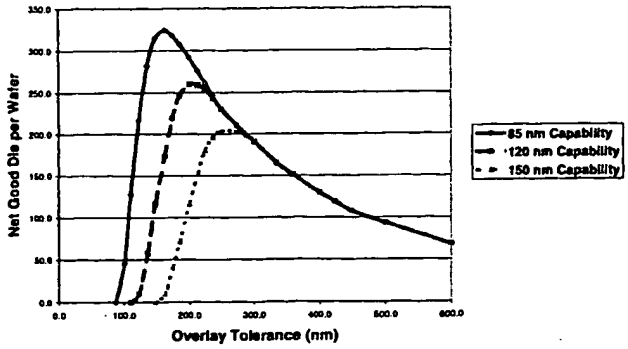


Figure 1: Net good die per wafer vs. overlay tolerance and capability, from ref. 2.

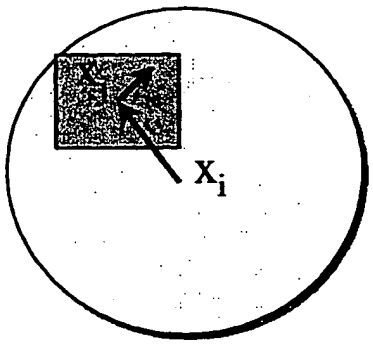


Figure 2: Definition of coordinate system showing point j within field i

Unfortunately, fab overlay capability is not a single fixed number that can be measured easily. Overlay performance varies over time and is highly dependent on the state of each exposure tool used in the process. Random variations in the systematic variables that govern overlay on each tool, and changes in other process parameters (e.g., film thickness, stress, CMP process conditions) can cause significant short term fluctuations in overlay capability. In addition, the close coupling between overlay and CD means that an overlay value which produces a good die under some conditions may result in a failing die if the CD varies slightly over time. For these and other related reasons, we can never set the overlay tolerance exactly at the peak of the net good die curve. If we did, daily variations would result in unacceptably high yield losses. We do, however, want to operate as closely as possible to the peak. This requires careful study of the overlay to yield correlation over an extended period of time to determine exactly how far the tolerance can be reduced without risking a major yield crash.

The question of yield to overlay correlation takes on even more importance as many fabs are trying to deal with the recent economic downturn by shrinking their devices without buying new equipment. If this is done successfully, the payoff is obvious; failure, on the other hand, would be painful to contemplate. While improved overlay could be achieved by exposing more send ahead wafers and accepting a higher rework rate, this would have a highly negative impact on productivity. In order to determine if further shrinks are possible with a given tool set, it is vital to understand the true relationship between overlay and yield.

It should be noted that while we speak of overlay as a single parameter, there are in fact many critical layers where overlay could cause a die to fail. There could even be second order failure modes where the overlay errors between layers 1 and 2 are within spec, as are the overlay errors between layers 2 and 3, but the unmeasured resultant errors between layers 1 and 3 could cause a failure. We will not address this special case in this paper. Instead, we will assume that a limited number of critical overlay pairings are known to be of the highest importance in determining yield, and we will focus on these critical layer pairs. In addition, overlay errors can take on both positive and negative values. Due to design issues, the threshold for overlay failure could actually be different for positive vs. negative offsets. Finally, there are overlay errors in both the x and y axes; the thresholds can be different for these two orientations. For simplicity, we will consider overlay errors in only one axis, and study only the absolute magnitude of the overlay errors. Thus, there is only one threshold to consider, and one overlay parameter to analyze. The analysis presented here can easily be extended to two axes and to asymmetric cases.

2. PROBLEMS CORRELATING OVERLAY TO YIELD

Traditional attempts at overlay to yield correlations have run into a number of obstacles. First, product wafers do not provide much overlay data to work with. Typical sample plans measure only 4 or 5 points per field on a very limited (5 to 10) number of fields. Only one or two wafers are measured, and often not on every lot. The true nature of lot to lot, wafer to wafer, field to field and site to site variations are impossible to comprehend from such limited data. As a result, most analyses rely on wafer or lot level averages. They plot the average yield for a given sample (wafer or lot) vs. the average or worst case overlay measured on that sample. With the rare exception of a completely misaligned sample, these analyses rarely show any meaningful correlation. By the very nature of the overlay equations[3,4], it is unlikely that the true worst case point will be

captured by such a limited set of measurements. Even if the worst point is actually measured, it still does not tell us what percentage of the wafer has high overlay errors. Depending on the values of the systematic fitting parameters, a few large errors could mean that every die on the wafer fails the overlay spec; that only a few die fail the spec; or, as is usually the case, that a fraction of the die pass and the rest fail. Given this lack of adequate input data, it is not surprising that overlay to yield analyses based on sample averages provide inadequate correlation to be meaningful.

The problem is compounded because overlay is but one of many factors which influence yield. As a result, wafer or lot level averages do not correlate well because the signal we are looking for is fairly small. If overlay were a major yield killer, it would be fairly obvious from failure analysis and visual observation in the fab. What we are trying to determine is what small fraction of the non-functional chips died because of overlay problems, and what fraction failed due to other sources of yield loss. We are looking for a small component of a larger number which is itself - hopefully - much less than one. The measured overlay numbers should also not be varying excessively. Thus we are trying to correlate a small number against a slowly varying input parameter in the presence of a larger and more variable background signal. The signal to noise ratio of the overlay induced yield loss compared to all other sources of yield loss is too small to be detected by sample level averaging.

A further complication is that the spatial signatures of an overlay problem can easily be mistaken for the signature of other systematic problems, and vice versa. In general, we decompose the sources of yield loss into two major categories, random and systematic defectivity[5]. Random losses include particles, foreign material and other non-repeating sources of loss. Systematic defectivity is defined as any non-random pattern of yield loss due to process, equipment or design issues. If we look at only a few wafers, all of the failing die may appear to be randomly distributed across the wafer. However, when we stack up hundreds of wafers, clear patterns emerge that help identify the systematic sources of yield loss. A common example is radial yield loss where the edge of the wafer has, on average, a lower yield rate than the center of the wafer. Numerous variations of these spatial patterns are seen in many processes; top to bottom or left to right variations in yield are not uncommon, and other, more complex patterns are also seen.

All of these spatial yield patterns complicate the problem of extracting the overlay induced yield losses. Overlay errors generally become large the further out the die is located on the wafer since scaling and rotational terms are linear in distance from the wafer center. Imagine if we were to look at a wafer with a pronounced spatial yield signature (Fig. 3a), measure the overlay error at many points (Fig. 3b) and correlate the results (Fig. 3c). The resulting correlation curve is grossly misleading. While a least squares regression shows a rough linear fit as indicated by the solid line, there is no way of knowing if this slope is due to a true causal relationship or if it is purely coincidental that both yield and overlay vary in a radial pattern but for different underlying reasons. In fact, one would expect an overlay induced yield loss to have little impact until a threshold is reached, and then to have a sudden, large effect on yield. The gradual slope of this curve strongly suggests that the seeming correlation between overlay and yield is really coincidental. There is nothing in this graph that tells us that a specific value of overlay error corresponds to a threshold where it becomes a yield limiting factor.

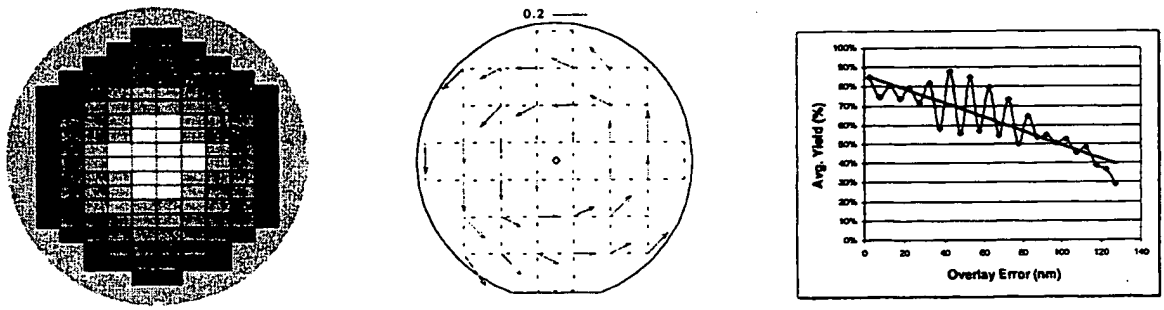


Figure 3: a) Wafer yield map showing a radial yield loss pattern. Darker colors indicate lower yield. b) Vector map of overlay errors for the same wafer with a rotational error. c) Apparent correlation between overlay and yield for this idealized case.

In summary, the problem of correlating overlay to yield is complicated by three main factors:

1) Wafer or lot level averages provide inadequate data
2) The signal to noise ratio is low because overlay is but one of many factors which influence yield
3) Spatial signatures of overlay vs. yield data can be confounded with other systematic sources of yield variation

These factors make it difficult to understand the true fab capability and to predict what further shrinks are possible. The poorly understood correlation may be forcing fabs to sandbag their overlay tolerances and accept less good die per wafer as a result. It also makes it difficult to detect subtle yield losses which could make the difference between a fab that turns a profit and a fab that is in danger of being shut down. In the following sections, we will describe how die level correlation coupled with some additional statistical procedures can provide a solution which addresses the first two issues, and shows promise for resolving at least part of the third obstacle to achieving useful overlay to yield correlations.

3. DIE LEVEL CORRELATION: MATHEMATICAL FRAMEWORK

Given that sample level analyses do not provide adequate correlation, the next logical step is to correlate the yield of each individual die with an estimate of the worst case overlay error within that specific die. Die level analysis is extremely rare in yield studies. Although probe data can be accessed by die, almost all other forms of parametric data, especially in-line data, are available only in summary form from a limited number of sites. Fortunately, the unique nature of overlay errors will allow us to make reasonable estimates of the die by die overlay values without increasing the amount of measured data. The overlay error at any site *j* within a field *i* (Fig. 2) can be written as the sum of the interfield and intrafield errors:

$$\Delta x_{ij} = \Delta x(\text{field i}) + \Delta x(\text{site j}) + \text{residual errors} \tag{1}$$

where the residuals themselves can be further broken down into interfield, intrafield, and random components. A simple example of the equations containing only the correctable parameters and residual errors is given by:

$$\Delta x_{ij} = \Delta x_w + S_x \bullet x_i - \theta_w \bullet y_i + M \bullet x_j + \theta_R \bullet y_j + R_i + R_j + R_{ij} \tag{2}$$

where x_w is the wafer translation in the x-direction
S_x is the wafer scaling error along the x-axis
θ_w is the wafer rotation error
M is the isotropic reticle magnification error
θ_R is the reticle rotation error
R_i is the residual error for the entire field
R_j is the residual error for site j in every field, and
R_{ij} is the residual component specific to site j within a specific field i.

Similar equations can be written for overlay errors in the y direction, Δy_{ij}.

A reasonable approximation of the fitting parameters can be obtained from the data measured in a typical production sampling plan[6]. Given these parameters and a set of equations appropriate to the exposure tool in question, we can make a first order estimate of the overlay errors for every die on the wafer. The equations can be evaluated quickly at many points within each die to determine the worst case error. A little knowledge of the residual errors could improve the estimates even further. For most leading edge exposure tools, the intrafield residuals R_j are the dominant term, being caused by mismatches between the lenses used to expose the different layers on the wafer. A comprehensive stepper matching program[7] could be utilized to create look up tables of these residual errors and further refine the die by die estimates. Look up tables could also be used to estimate the interfield residuals, though in general these are fairly small. The random residual terms R_{ij} are due to all other sources of uncorrectable wafer distortions- chucking errors, non-linear wafer expansion, localized stress and metrology errors. In general these terms are much smaller than the other systematic and residual errors and may be ignored.

The estimated die by die overlay errors can now be correlated to the die yield. Since yield is a dichotomous response with only two possible values – pass or fail – this data is not amenable to linear regression. Fig. 3 shows a modeled example of this correlation; the y-axis data is either a zero (fail) or a one (pass). Many values of overlay can have more than one yield value, since a given worst case overlay error can occur on different die. In many cases, a given overlay error can have both good and bad die. This is not an uncommon result in pass/fail situations; much of the statistical machinery used in medical research has been tailored explicitly for such problems.

There is, in fact, a great deal of similarity between the overlay-yield question and the classic medical question of establishing lethal doses of various toxins. While the consequences of producing a bad die are rarely fatal, both problems involve non-repeatable experiments, noisy data, and numerous additional variables and complicating factors. In both cases, it is nearly

impossible to say that a value below some limit is always good and a value above the limit is always bad. The goal is to establish a threshold that is likely to cause a problem; in medical terminology, this value is known as LD_{50}, a statistical estimate of the dose that would probably be lethal in 50% of cases. For obvious reasons in both the medical and IC cases, it is desirable to establish a good estimate of this parameter with the minimum amount of excess experimentation. Although 50% may not be the best value for yield issues, it is our goal to establish a similar statistical metric of the correlation between overlay and yield in IC manufacturing. To demonstrate how this formalism can be used to analyze overlay data, we present the following example.

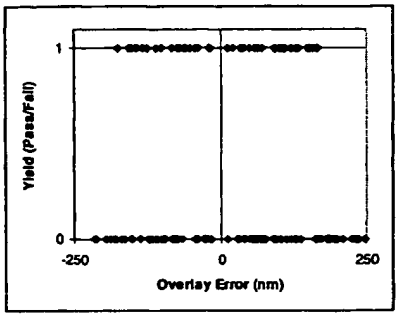


Figure 4: Die by die yield as a function of the modeled overlay error at each die.

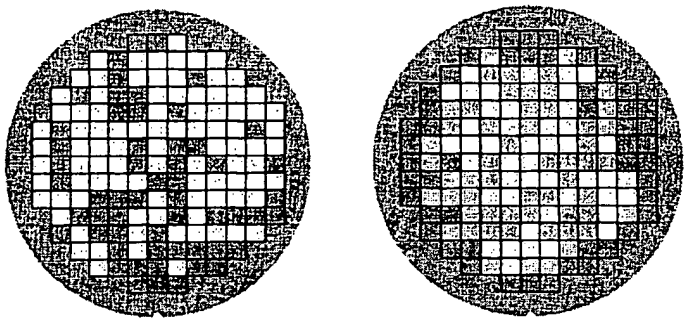


Figure 5: Simulated yield maps. a) Single wafer map; light die are good, dark die fail. b) Average of 50 wafers. Darker colors indicate lower yield.

4. DIE BY DIE CORRELATION: SIMULATED RESULTS

In order to test the die by die algorithm, we generated simulated yield and overlay results. While this does not guarantee that the formalism will work with real fab data, it has the advantage that we can manually tweak the various sources of yield loss and study the resulting overlay-yield response in the presence of controlled variations. We modeled the yield at each die as

$$Y = Y_r * Y_s * Y_o \tag{3}$$

where Y_r is the random yield term, Y_s the systematic yield term, and Y_o the overlay induced yield term. All three parameters can take on values of 0 or 1; the die will yield only if all three terms equal one. The random yield value at each die was set using numbers randomly selected from a Poisson distribution. The systematic yield term was set to kill an average number of the die, with the losses being selectively higher for die near the outer edge of the wafer. The overlay term was modeled by calculating the systematic overlay error at all four corners of the die and adding in a small random residual error at each site. The largest of these four values was then compared to a threshold value and Y_o was set to zero if the error exceeded the threshold. To simulate real world effects such as CD variation, the threshold included a random variation about the target value. The option exists to model both x and y overlay and set Y_o equal zero if either term exceeds a threshold value.

Each time the program is run, the user inputs the baseline systematic overlay parameters (translation, rotation, scale, reticle rotation and magnification), the random defect density and the systematic yield loss percentage. Additional parameters may be set to vary the magnitude of the variation in the overlay threshold, the radial dependence of the systematic yield variation, and the range of residual overlay errors. These parameters allow us to simulate conditions which cover the range from high yielding wafers to very low yielding wafers, with varying mixes of random, systematic and overlay induced yield loss. The model generates wafers with 149 die per wafer and computes a worst case overlay value and a yield value for each die. The program varies the overlay fitting parameters by small random amounts for each wafer to simulate wafer alignment errors. We typically compute 5 wafers at a time, adjust the input parameters slightly to mimic lot to lot variation, and repeat the process a number of times to generate a large group of wafers.

A typical example is shown in Figures 4 to 7. The model parameters were 0.3 defects/sq. cm. for the Poisson defect distribution, 11% average systematic yield loss (the range was 5-17%), and a 200 nm overlay threshold. The overlay fitting parameters were all small except for a 3 ppm wafer rotation error; the intrafield residuals were normally distributed about 0 with a 3 sigma of 30 nm. Fig. 5a shows the yield pattern for a single wafer, while Fig. 5b shows the average yield for 50

wafers. Figs. 6a through 6c show the systematic, random and overlay induced losses averaged over 50 wafers. Note that the radial yield loss pattern due to systematic and overlay problems is not apparent in any individual wafer plot; the loss of a few die near the edge is no more pronounced than the loss of die elsewhere due to random defects. However, the pattern becomes quite evident when a large number of wafers is stacked up and averaged. Fig. 7 shows the wafer level averages of overlay vs. yield; Fig. 7a shows the average overlay value on each wafer vs. yield, while Fig. 7b shows the range in overlay measurements which is sometimes tracked to detect excursions in the overlay fitting terms. Although a slight linear relation exists between the increased range and reduced yield, the goodness of fit parameters are extremely small and indicate little statistical significance to the fit.

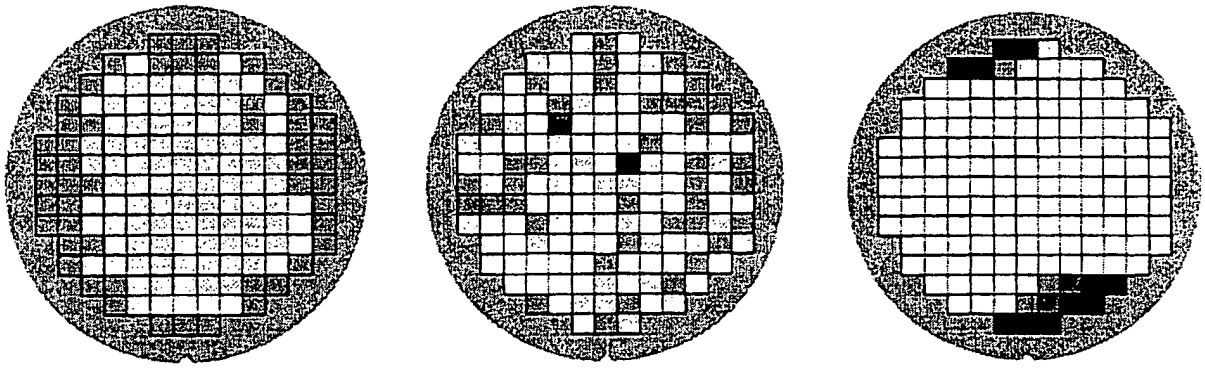


Figure 6: Components of yield loss, averaged over 50 wafers. a) Systematic yield loss, b) Random yield loss, c) Overlay induced yield loss. The product of these three terms gives the total yield shown earlier in figure 5b.

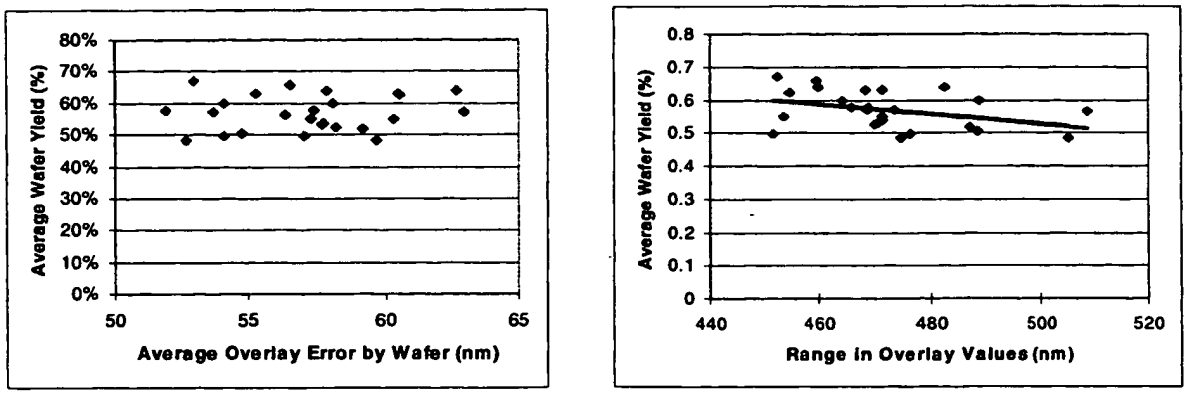


Figure 7: a) Average wafer level yield vs. average wafer level overlay error and b) vs. the range in overlay values across the wafer.

The die by die correlations, on the other hand, show the effects we had anticipated. Fig. 4, discussed earlier, shows the raw correlation. At first glance, it is not obvious that there is a correlation, although closer inspection shows that there are no good die with overlay errors greater than 200 nm. To add statistical rigor to this observation, we binned the overlay data in 20 nm increments and calculated summary statistics for each interval. Figure 8a shows a graph of mean yield by overlay bin interval. It can be seen that the yield declines rapidly for the overlay intervals above 200nm. A chi-squared test was performed on yield by overlay interval for the complete data set, and was found to be significant at > 99.9% confidence. We therefore reject the null hypothesis that yield is independent of overlay. Each interval was then compared to the control interval by generating contingency tables and performing chi-squared tests. Table 1 shows a typical contingency table for this chi-squared analysis. The number in each cell represents frequency e.g. there were 217 good die with low overlay error.

This is an ideal method for analyzing a binary variable such as die level yield. In this way, the P value for the null hypothesis that yield was independent of overlay was estimated for each 20nm overlay interval. Fig. 8b shows that the chi-squared

values are not significant for overlay intervals below 200nm, but for intervals above this the P-values are <0.001, indicating with high confidence that die with overlay errors in the 190-210 nm range and above yield differently than die in the -10 to +10 nm range. This gives us a reasonably accurate estimate of the 'real' overlay tolerance of a given product and allows us to estimate the yield impact based on the current overlay capability.

From a statistical perspective, caution should be exercised when performing multiple comparisons of this kind because the overall type 1 (alpha) error rate is inflated as the number of tests increases. i.e. the chances of falsely rejecting the null hypothesis are increased. This problem is well document in the statistical literature[8]. To minimize the alpha risk, we can use the data shown in Fig. 8a to give us an indication of which intervals are of the most interest. In this case, the yield roll-off occurs somewhere between 180 and 220 nm, so performing a smaller number of chi-squared tests around these intervals would be safer and more statistically sound.

	Low Overlay Error -10 to +10 nm	High Overlay Error 20 to 30 nm
Pass	217	874
Fail	89	387

Table 1: Sample contingency table input comparing a test range to the control range.

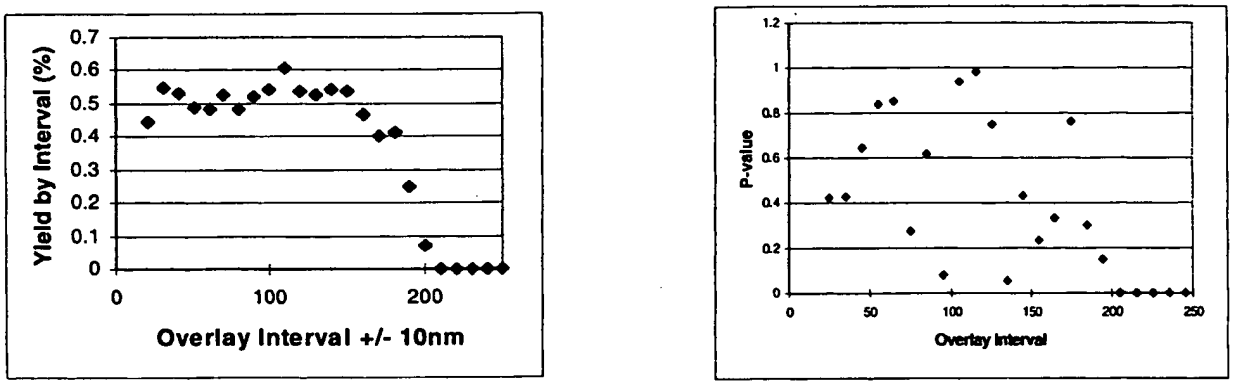


Figure. 8: a) Mean yield and b) P values from chi-squared tests vs. overlay bin interval

5. REGRESSION ANALYSIS

Linear regression can not be used for die level analysis since the response is dichotomous (pass/fail) instead of continuous. Logistic regression using the logit function is ideally suited for this case[9], and we will use this approach to estimate the probability of a die yielding as a function of the overlay error for that die. The model used is

$$\text{Logit}(p) = \text{constant} + A * \text{overlay error} \quad (4)$$

where $\text{logit}(p) = \text{Ln}(p/(1-p))$
p = probability of an event, i.e., the probability that the die is good, and
A is a fitting coefficient we will determine from the analysis.

Higher order terms can also be added to the equation but we felt initially that there was no a priori justification for anything beyond the simple linear term. At first glance it might appear that this equation could not fit a very sharp threshold for overlay induced failure. Tests on sample inputs with an absolute hard cutoff at 200 nm and no other sources of yield loss – in other words, a perfect square wave response – were actually fit quite well by this model.

When we study overlay induced yield losses alone, this method works very well. Fig. 9 shows the computed logit curve; the 50% yield point is exactly at our preset threshold, which gives us the overlay equivalent of LD_{50}. This threshold is the point

where 50% of the die with an overlay error of this magnitude will fail. When we add in random yield losses, the curves are not as sharp, but the 50% line still hits the assumed threshold of 200 nm (Fig. 9b). However, when we add in systematic yield loss, the curves shift, and the 50% threshold moves from the simulated limit. We know that the value of 200 nm should have some significance since our simulator used that number as the baseline threshold value for overlay induced yield failure; without that prior knowledge, there is no way to extract a threshold value from this fit. Logistic regression assumes an S-shaped relationship between yield and the variable parameter, with a range from zero to one. It is possible that with the addition of large systematic losses, we have distorted the curve so badly that the fit is no longer obtainable. In addition, we no longer have a control "dose" where yield is always guaranteed to be one; this may be the source of the problem. We will pursue additional modifications of the logistic regression procedure, including the use of higher order terms, in an effort to find a model that yields a rational overlay threshold in the presence of multiple sources of yield loss.

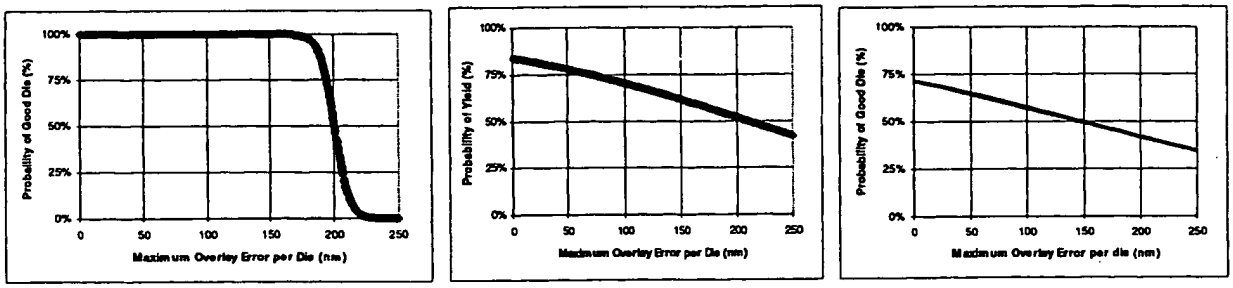


Figure 9: Logistic regression results. a) Overlay induced losses only, b) overlay and random yield losses, c) overlay, random and systematic yield losses

6. CONCLUSIONS

Our analysis has shown that die to die correlation, as opposed to wafer or lot level averages, is a powerful tool for extracting the overlay induced yield losses from the other sources of random and systematic yield loss. The equations describing overlay as a function of location have been studied intensively for many years. This analysis takes advantage of these well established models to generate the equivalent of a large amount of overlay data without increasing the number of measurements required. This is a unique advantage particular to overlay analysis which should certainly be exploited.

The technique is far from perfect; if the systematic loss pattern matches the spatial signature of the overlay errors, there will be no way to separate these two effects. In most cases, however, if the defect and other non-overlay related losses are small, the amount of wafer to wafer variation will probably be sufficient to allow some measure of the overlay induced losses to be extracted from the data. While the averages of the overlay and other spatial signatures may appear similar, the fluctuations between individual wafers – measured over a large number of wafers – should prevent the overlay signal from being completely obscured. It remains to be seen from real data if this analysis will hold up in real world situations.

Logistic regression may hold some promise for further improving this analysis. Since it appears that systematic losses create the most problems for this method, we intend to develop an algorithm to separate random and systematic losses, then retry the regression on the corrected data. It is also possible that we could reject die where the analysis shows that random defectivity was the clear cause of yield loss. (A failing die with a low overlay error surrounded by good die with higher overlay errors would be a clear example of such a flier). Regardless of whether or not these proposed additions prove to be successful, it is clear that die by die correlation should always provide a better fit to the data than sample level averaging.

ACKNOWLEDGMENTS

The authors wish to thank Linda Buck, Buffy and Alisa Preil for their help in preparing this manuscript.

REFERENCES

[1] *National Technology Roadmap for Semiconductors*, 1994 and 1997 editions published by Sematech, and unpublished data from the preliminary 1999 revision.

[2] W. H. Arnold and J. Greeneich, "Impact of Stepper Overlay on Advanced Design Rules", Proceedings of the OCG Microlithography Seminar, pp. 87-105, (1993).

[3] J. D. Armitage and J. P. Kirk, "Analysis of Overlay Distortion Patterns", SPIE vol. 921, p. 207-222 (1988) and references therein.

[4] M. A. van den Brink, C.G.M. de Mol and R.A. George, "Matching Performance for Multiple Wafer Steppers Using an Advanced Metrology Procedure", SPIE vol. 921, p. 180-207 (1988).

[5] A. Y. Wong, "Statistical Micro Yield Modeling", Semiconductor International, Nov. 1996, pp. 139-148.

[6] I. Fink, N. Sullivan and J.S. Lekas, "Overlay Sample Plan Optimization for the Detection of Higher Order Contributions to Misalignment", SPIE vol. 2196, pp. 389-399 (1994).

[7] M. E Preil, "A Comprehensive Approach to Overlay Improvement", KLA-Tencor Yield Management Consulting methodologies, unpublished.

[8] John A. Rice, "*Mathematical Statistics and Data Analysis*", 2nd edition, Duxberry Press (1995).

[9] P. McCullagh and J .A Nelder, "Generalized Linear Models", 2nd edition, Chapman and Hall, London, (1989).

Assessment of thermal loading-induced distortions in optical photomasks due to *e*-beam multipass patterning

Bassam Shamoun[a] and Roxann Engelstad
University of Wisconsin—Madison, Madison, Wisconsin 53706

David Trost
Etec Systems, Inc., Hayward, California 94545

(Received 7 August 1998; accepted 14 September 1998)

Thermal loading-induced distortion in the photomask during *e*-beam patterning has recently received special attention due to its significant contribution to overlay errors. Multipass *e*-beam writing, a strategy proposed to reduce the heating effects and associated distortions, was simulated using three-dimensional finite element models. Thermal responses of the photomask during multipass patterning were determined and global in-plane distortions were calculated. For the given system exposure conditions of 40 $\mu C/cm^2$ at 50 keV, the average value of the 3σ pattern placement error due to the bulk heating of the photomask obtained from multipass writing was found to be $\approx$3.5 nm which is 28% lower than that of single pass writing. Parametric studies showed that thermal radiation has a large influence on the mask cooling. © *1998 American Vacuum Society.* [S0734-211X(98)06506-8]

I. INTRODUCTION

The technological demand for higher capacity memory chips has created a special challenge for photomask makers to meet the requirements of pattern placement accuracy and throughput for advanced masks. Thermal loading-induced distortions caused by resist substrate heating due to the *e*-beam energy deposition during mask patterning is a major contributor to pattern placement errors. Given the system exposure conditions and mask material properties, previous investigations showed that pattern placement accuracy is significantly affected by local[1–3] and global[4] heating of the mask substrate. One of the suggested methods to reduce the heating effects of the photomask during the patterning process is to use the multipass writing strategy in which the required dose to develop the resist is applied through a number of passes across the substrate. The work that has been done in the past focused on the effects of local heating on resist sensitivity during *e*-beam exposure. Studies to examine the effects of global heating on the pattern placement accuracy, however, have just begun. Due to the difficulty of direct measurement of temperature rise in the resist substrate, theoretical models[5–7] have been proposed to estimate the heating effects during mask fabrication. Computer simulations of the patterning process provide valuable tools for the assessment of the *e*-beam patterning induced distortions. Data can subsequently be used to adjust system parameters leading to process optimization and minimization of pattern placement errors. In this article simulation results of thermal-induced distortions in the photomask during multipass *e*-beam writing obtained from the three-dimensional finite element (FE) analyses are presented. Comparison with the thermal in-plane distortions (IPD) obtained from single pass writing will also be made.

[a]Electronic mail: bshamoun@etec.com

II. FINITE ELEMENT MODEL

The FE models simulating the thermal IPD used a 6 in.×6 in. Format optical reticle with a pattern area of 132 mm×132 mm. Figure 1 shows a schematic diagram of the system geometry employed in the analysis. The mask model consisted of three layers: Novolac-based resist (0.4 μm), chrome (0.08 μm), and quartz substrate (6.4 mm). The original stress for the resist and the chrome layers is taken to be 10 and 25 MPa, respectively. Thermal properties are listed in Table I. The patterning area of the mask was divided into fields, with each field subdivided into a specified number of finite elements. The FE models incorporate eight-node isoparametric brick elements for the mask substrate and four-node elastic shell elements for the chrome and resist layers. In the simulation, a thermomechanical load is applied to each field in the pattern area of the mask in a sequential manner through a specified number of load steps following the serpentine writing style illustrated in Fig. 2. By tracking the displacement of each node in the model for several load steps the final IPD map is obtained. The finite element software ANSYS[8] was used to perform all calculations.

The total *e*-beam energy deposited in the mask substrate is determined from the beam parameters (full dose of 40 $\mu C/cm^2$ at 50 keV). The fractional energy deposited in the mask is based on Monte Carlo simulations. As a representative worst case scenario the pattern density was assumed to be 100% in the calculations. Assuming the mask is placed in a vacuum chamber, the thermal model considered conduction and radiation heat transfer mechanisms, and used adiabatic boundary conditions (i.e., no heat flow) at the edges. Equivalent surface heat flux and volumetric heat generation were used to simulate the thermal energy deposition in the mask. Support conditions for the reticle consisted of a "2-2-2" kinematic mount located at a radial distance of 28.7 mm from the center of the substrate. Figure 1 indicates the de-

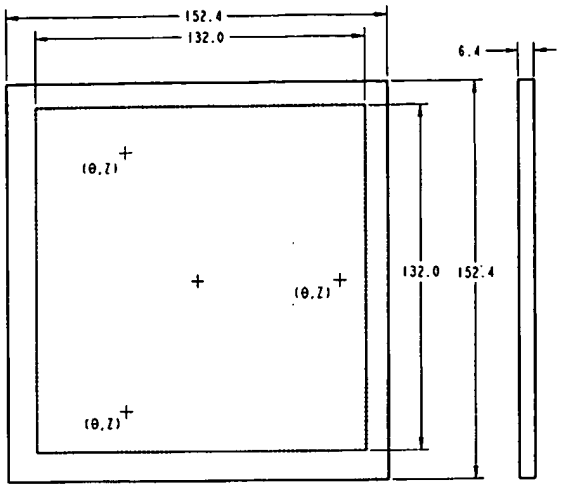


FIG. 1. Schematic diagram of the 6 in.×6 in. optical reticle; all dimensions are in mm. Location of a traditional "2-2-2" kinematic mount is shown with the translational degrees of freedom constrained at each mounting point given in parentheses.

grees of freedom constrained at each mounting point. A displacement vector of the central node of each exposed field in the pattern area was used as a measure of the resulting distortion of that field from its original position at the end of the patterning process. The magnitude and the direction of each displacement vector are a function of several parameters such as the patterning time, writing style, amount of the absorbed dose, and thermal and mechanical boundary conditions.

III. SIMULATION RESULTS

The patterning process in this analysis was simulated using four passes with no waiting period between successive passes. The *e*-beam energy deposited in the mask substrate for each writing pass was assumed to be one fourth of the full dose required to develop the resist. The patterning time per pass was accordingly taken to be one fourth of the total time (6 h) required to pattern the mask in a single pass writing scheme. As a result, the maximum temperature rise (ΔT_{max}) at the exposed surface of the mask in the multipass writing was expected to be lower than that obtained in the single pass writing. It is important to note, however, that the temperature distribution obtained at the end of each writing pass was used as initial conditions for the subsequent pass.

Both the average temperature rise (ΔT_{ave}) and ΔT_{max} of the mask were plotted as a function of the patterning time for all four passes. Figure 3 shows that the average temperature

TABLE I. Thermal properties of reticle materials.

	Conductivity (W/m K)	Specific heat (J/kg K)	Thermal expansion coefficient (ppm/K)
Resist	0.20	1500	83.0
Chrome	93.7	4605	4.20
Quartz	1.46	750	0.55

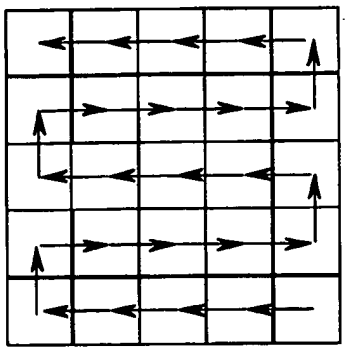


FIG. 2. Schematic diagram showing the serpentine writing style used for photomask *e*-beam patterning.

of the mask continued to rise at the beginning of the patterning time in the first pass. As the radiation losses increased, the rate of increase of ΔT_{ave} slowed down reaching a quasi-steady state in about 1.2 h. This behavior is similar to that observed in the simulation of the single pass patterning.[4] While ΔT_{ave} achieved a steady state in a relatively short time, ΔT_{max} continued to oscillate between two bounds. The oscillatory behavior of ΔT_{max} is mainly due to the writing style (serpentine in this case) and the applied adiabatic boundary conditions. The effect of the boundary conditions appears mostly in the region near the edges and the corners of the pattern area where the temperature rise tends to jump, forming a peak. As the beam moves toward the center and away from the boundaries the temperature rise drops. It can be seen that at the end of each writing pass and at the beginning of a new pass ΔT_{max} experienced a sudden drop (a fraction of a degree) as the beam moved from the top to the bottom corner of the pattern area as shown in Fig. 3. In general, the average value of ΔT_{max} obtained during the multipass writing was found to be 36% lower than that from the single pass writing.

Figure 4 shows examples of the contour plots of the temperature distribution following the first and the last load steps of the first and second writing passes obtained from the FE

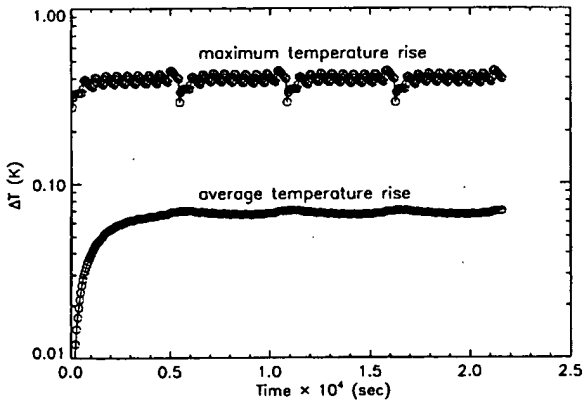


FIG. 3. Average and maximum temperature rise in the photomask due to the global heating as a function of time obtained from the FE simulation of the multipass *e*-beam patterning.

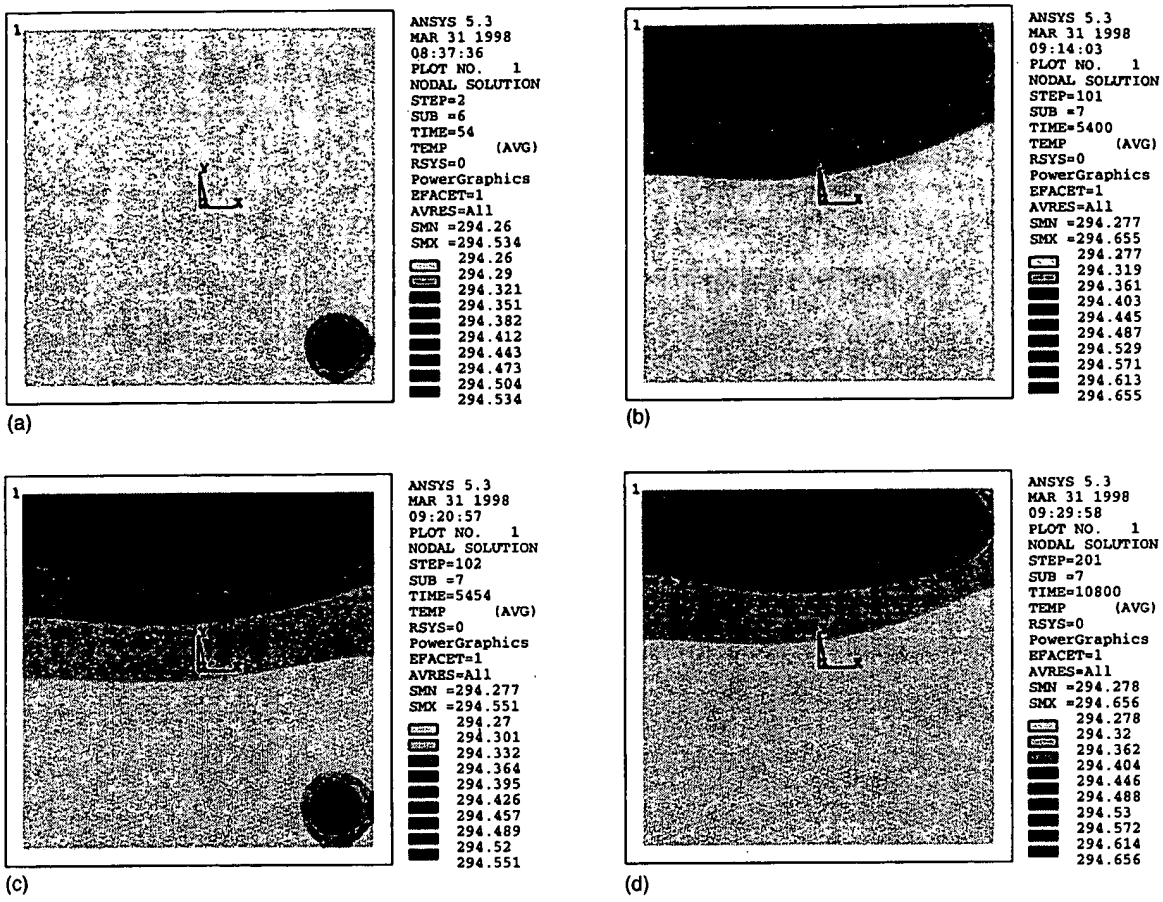


FIG. 4. Results of the FE thermal simulation of the *e*-beam multipass patterning for the optical mask. Contour plots showing examples of the temperature distribution following (a) first load step in the first pass; $\Delta T_{max}=0.274$ K, (b) last load step in the first pass; $\Delta T_{max}=0.395$ K, (c) first load step in the second pass; $\Delta T_{max}=0.291$ K, and (d) last load step in the second pass; $\Delta T_{max}=0.396$ K. The initial temperature of the mask was taken to be 294.26 K. FE scale in Kelvin.

analyses of the multipass writing. It should be noted that the patterning in each writing pass began at the lower right corner of the substrate. It is clear that the temperature distribution of the mask at the end of each writing pass is not uniform. This is mainly due to the poor thermal conductivity of the quartz substrate, which prevented the mask from reaching a thermal equilibrium state at a time constant shorter than the exposure time.

Using the data from the FE thermal analysis, the global IPDs were calculated at the end of each writing pass. The distortion maps obtained for the four passes are shown in Fig. 5. The displacement vectors are shown to be radially oriented outward. The mounting conditions of mask structures are believed to have a significant effect on the distribution of the IPD vectors. From a closer look at the IPD map from the first pass [shown in Fig. 5(a)], one should note that the lower right corner field in the pattern area has zero displacement. The reason for this is the assumption made in which the patterns in this field (initially at room temperature) are written on the undistorted pattern area. As the patterning process continues, however, the first field as well as the subsequent fields begins to distort.

The 3σ pattern placement error was also calculated for each pass and found to be 3.75 nm for the first pass and 3.5 nm for subsequent passes. It is interesting to note that the maximum value of the 3σ pattern placement error occurred in the first pass. This, however, is expected since the mask in the first pass was still cold and the expansion due to the temperature rise is highly nonuniform causing the displacement vectors to be larger in magnitude. As the average temperature of the mask increases, the exposed area gains more freedom to move in other directions, therefore reducing the magnitude of the displacement vectors somewhat. Again, in comparison with the single pass writing,[4] the 3σ pattern placement error obtained from the multipass writing was 28% lower. In general, the resulting IPD maps indicate that there is symmetry in the distribution of the displacement vectors. Unless the pattern density distribution in the patterning area of the actual mask is very different from one substrate to another correction for the pattern placement error can be made directly during fabrication. However, the effect of pattern density distribution on the resulting IPD needs to be further investigated.

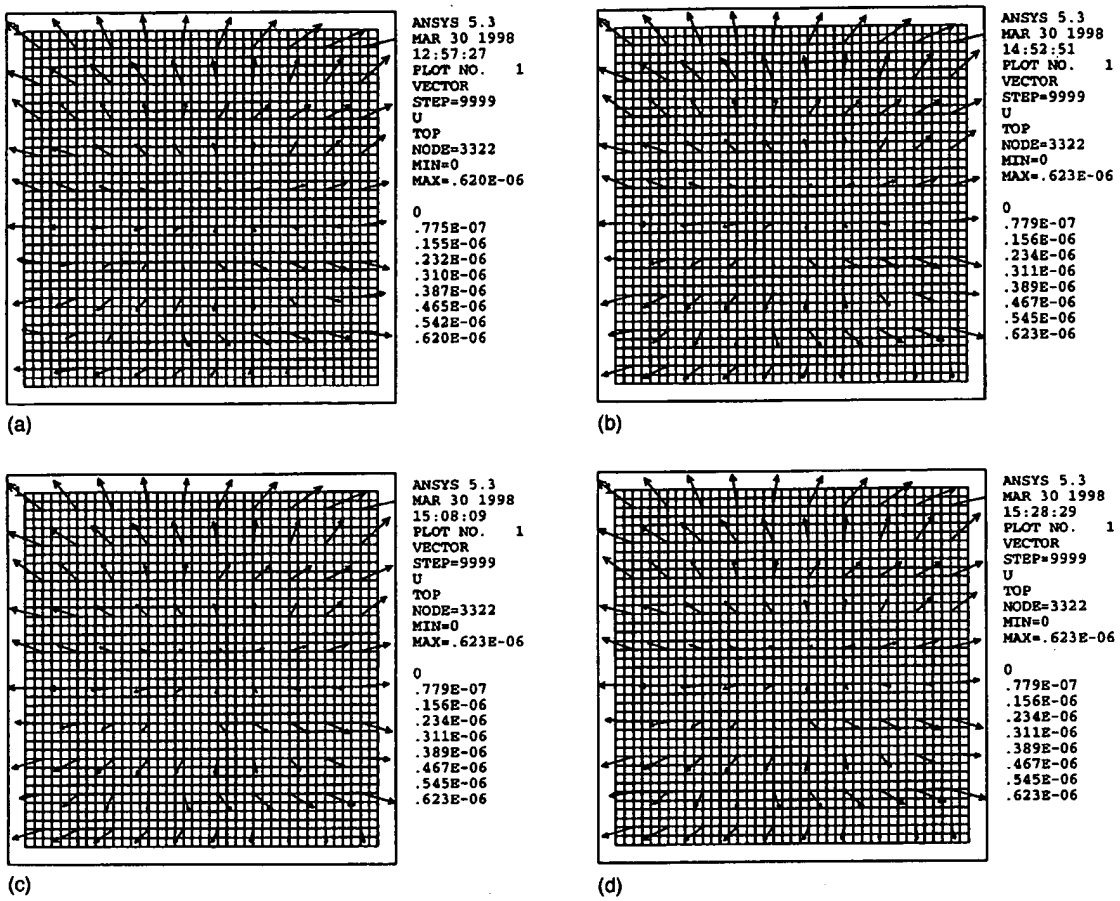


FIG. 5. Thermal loading IPD maps obtained from the FE analysis of the multipass *e*-beam patterning for the optical mask (a) thermal IPD following the first pass; Max.=6.20 nm, 3σ=3.75 nm, (b) thermal IPD following the second pass; Max.=6.23 nm, 3σ=3.5 nm, (c) thermal IPD following the third pass; Max.=6.23 nm, σ=3.5 nm, and (d) thermal IPD following the fourth pass; Max.=6.23 nm, 3σ=3.5 nm. FE scales in cm.

IV. EFFECT OF THERMAL RADIATION

In any enclosure, radiation may experience multiple reflections of all surfaces, with partial absorption occurring at each. The previous results obtained from the FE analysis are limited by the assumption of blackbody radiation. In the actual *e*-beam writing apparatus this may not be the case since the surroundings are likely to reflect some of the emitted radiation back onto the mask surface causing a further increase in its temperature. Although the amount of the reflected radiation depends on the geometrical configuration of the system components and the optical properties of the reflecting materials, it is useful to consider the worst case scenario by assuming the case of a perfect reflector. That is, no heat is dissipated from the surface of the mask by radiation, and conduction is the only mechanism for heat transfer within the mask substrate during *e*-beam exposure. This is also the case in which the results obtained from the FE simulation can be compared with the analytical solution from a simplified form of the diffusion equation. Figure 6 shows the results of ΔT_{ave} and ΔT_{max} as a function of the patterning time for both single and multipass writing schemes. It can be seen that the temperature rise increases linearly with time and it is much higher than that shown in Fig. 3. The resulting IPDs for this case are shown in Table II. These results suggest that thermal radiation plays an important role in dissipating a large amount of heat from the mask, therefore, reducing the heating effects and the associated thermal

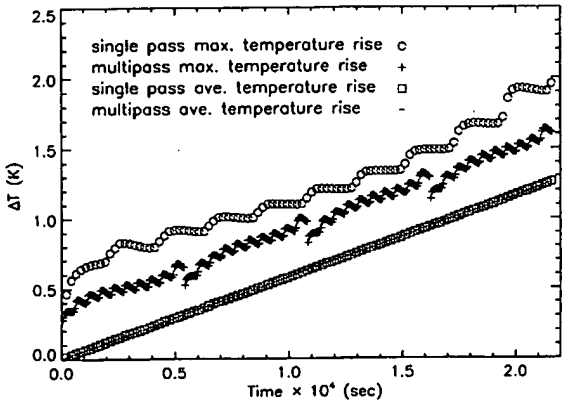


FIG. 6. Maximum and average temperature rise of the photomask due to global heating as a function of patterning time obtained from the FE analysis during single and multipass writing. Radiation cooling was not included in the model simulation.

TABLE II. Pattern placement error (3σ) induced in the photomask during the single and multipass *e*-beam patterning. Thermal radiation was not included in the FE simulation.

Pass No.	Single pass 3σ (nm)	Multipass 3σ (nm)
1	46.2	11.7
2	—	19.3
3	—	27.5
4	—	36.5

distortions. The geometrical features of the radiation exchange between the mask and the surroundings are an important factor for accurate assessment of heating-induced distortions in the photomask during *e*-beam patterning.

V. CONCLUSIONS

The nonuniform heating of the photomask substrate, the main cause of thermal distortions during *e*-beam patterning, can be minimized using a multipass writing strategy. The thermal-induced distortions in the photomask during patterning have been investigated using three-dimensional FE analysis. Global IPDs were calculated and found to be lower than those obtained from the single pass writing. The magnitude and direction of the distortion vectors obtained from the FE analysis were found to be dependent on the amount of absorbed dose, mounting conditions, and writing style. The FE simulations showed that thermal radiation is an effective means of cooling which keeps the average temperature rise within the desired range. In general, it is possible to correct these distortions for all pattern densities using these model simulations.

ACKNOWLEDGMENTS

This research has been supported in part by SEMATECH, the Semiconductor Research Corporation (SRC), and the National Science Foundation.

[1]S. Babin, J. Vac. Sci. Technol. B **15**, 2209 (1997).
[2]N. K. Ebib and R. J. Kvitek, J. Vac. Sci. Technol. B **7**, 1502 (1989).
[3]E. Kratschmer, J. Vac. Sci. Technol. B **8**, 1898 (1990).
[4]B. Shamoun, M. Sprague, R. Engelstad, and F. Cerrina, Proc. SPIE **3331**, 275 (1998).
[5]S. Babin, I. Yu. Kuzmin, and G. Sergeev, Proc. SPIE **3236**, 464 (1997).
[6]Z. Cui, Proc. SPIE **3331**, 420 (1998).
[7]T. R. Groves, J. Vac. Sci. Technol. B **14**, 3839 (1996).
[8]*ANSYS User's Manual* (Swanson Analysis Systems, Inc., 1996), Rev. 5.3.

Characterizing Overlay Registration of Concentric 5X and 1X Stepper Exposure Fields using Interfield Data

Frank Goodwin[a] and Joseph C. Pellegrini[b]

[a]AMI Semiconductors, Pocatello, ID 83201-2798
[b]New Vision Systems, Inc., Watertown, MA 02172

ABSTRACT

The cost advantages associated with implementing a mix-and-match photolithography process have led to a dramatic increase in the interest and development of these manufacturing environments. This is especially true for older fabs with high production lithography tools already in place but technology that has increased beyond the capability of the tools. For the process engineer the challenge is to define a method of optimizing the exposure field registration between each of the different imaging systems. In this paper a procedure used to evaluate intrafield and interfield overlay errors between six ASML 5X steppers and sixteen Ultratech 1X steppers is described. With this technique reticle data, stage registration and a commercially available software analysis package are used to model pattern displacement of each stepper within this population. Wafers from each stepper are first patterned with nine fields, each consisting of a 9 by 9 array of ASML alignment marks. The X and Y stage coordinate of each alignment mark is then measured using a standard ASML 5500/60 intrafield analysis routine. Spreadsheeting the resulting stage registration data, subtracting the expected or "ideal" stage position and correcting for any reticle pattern shifts, grid and intrafield data are obtained. Using this process a data sheet for each stepper was developed and, once formatted properly, loaded onto the software analysis package for registration modeling. Use of multiple exposure fields per wafer enabled the software to characterize both intrafield and interfield registration by first modeling the grid errors, subtracting these values, and then performing intrafield analysis on the remaining data. Further, by collapsing the intrafield data into a single field a "lens fingerprint" of each stepper lens was derived. Using vector subtraction a direct comparison was made between the lenses of each stepper and an indexed table of exposure field translation errors created. The stepper lenses were also sorted from best to worst matches. This approach generated the required 231 paired data sets needed to match each stepper to all others while exposing and measuring only 44 wafers (2 per stepper) and required no artifact wafers. Measured evaluation results will be reviewed and expansion of this procedure to mapping 1X wide field lenses and matching of non-concentric exposure fields discussed.

Keywords: mix-and-match, concentric matching, non-concentric matching, collapsed field, MONO-LITH, ASML, Ultratech, overlay registration

1. INTRODUCTION

Containment of device fabrication costs is fast becoming the dominant issue for semiconductor manufacturers looking to increase the integration of their devices[1]. With the cost of new high volume fabrication facilities reaching the billion dollar mark, manufacturers now look for alternative equipment schemes to control capital investment as well as upgrade technology[2,3]. As the lithography area represents a large investment in any facility, mix and match lithography environments have been successfully implemented as a cost savings measure[1]. In most cases, the less costly high throughput tools are used to pattern non-critical layers while the expensive high NA steppers image the critical levels. This approach allows older fabs to move to more aggressive design rules while maintaining production capacity and reducing initial investment. Increased technology, however, infringes upon device overlay budgets. So in order to maintain and improve device

0227-786X/97/$10.00

yields, registration errors for each exposure tool must be characterized and matched to each other. This paper reports on an evaluation of intrafield and interfield overlay errors between six ASML 5X and sixteen Ultratech wide field 1X steppers.

2. PROCESS DESCRIPTION

The equipment set used in this evaluation consisted of 12 Ultratech model 1000s, 4 Ultratech model 1100s, 2 ASML 2500/40s and 4 ASML 5500/60s. Concentric matching of a 14mm square field was performed as this is the field size used in our mix-and-match process and is the largest common among all tools in the equipment set, figure 1 (Ultratech usable lens area). The necessary interfield and intrafield data for concentric matching was collected by exposing 2 wafers on each stepper with a grid array of nine fields. Each pattern consisted of a 9 X 9 array of ASML primary alignment marks spaced 1.5 mm apart on the wafer, figure 2. A metrology job was then set up on the ASML 5500/60s that performed a local alignment to each of the marks and recorded the stage coordinates. This technique establishes this coordinate system as the absolute reference. After correcting for reticle errors, the overlay data was modeled using MONO-LITH®'s registration analysis software package and lens matching system, LEMSYS™.

Usually, characterization of stepper registration performance requires the use of artifact wafers. Overlay patterns are exposed and etched onto these wafers. Then, they are exposed a the second level of the registration pattern. Grid and intrafield data are obtained by measuring the localized displacement of the second layer image to the first. However, the procedure used in this test does not require an artifact wafer but references directly to the stage coordinate system of the ASML stepper. The ASML 5500/60 stage is configured with an interferometer mirror system used to measure stage X, Y, and Φ_Z parameters. These parameters are controlled by a 3 axis linear magnetic motor combination in an H configuration. The precision and reproducibility the ASML 5500/60 stage and the phase grating alignment system have been well documented[4,5]. For this evaluation the stage coordinate system was assumed to be close to ideal.

In our mix-and-match processes first level and all critical layers are imaged on the ASML steppers. The Ultratech steppers are then used to expose the non-critical levels. This alignment strategy uses the blind stepping precision of the ASML stepper to establish the stepping grid. The Ultratechs then use their field-by-field alignment to overlay to this grid. To mimic this, an array of nine sets of Ultratech horizontal alignment marks (HAMs) is first exposed on the test wafers using the ASML steppers and then developed. Each exposure field is then exposed on the Ultratechs after aligning to a set of the HAMs. Although this procedure does not provide any characterization of the Ultratech blind stepping precision it does allow evaluation of their field alignment accuracy.

Lens matching was performed on each stepper. Most stepper manufacturers monitor and match lens distortions of the systems delivered to their customers. But when employing different steppers, there is no guarantee that the distortion errors for lenses from one manufacturer match those for lenses of another manufacturer. To determine the lens differences of this stepper population, a collapsed field analysis was performed using each stepper's distortion data and LEMSYS™, MONO-LITH®'s lens matching software package. Collapsed field analysis removes grid errors from the registration data, merges data from all fields into a single field, and selects the median value of the vectors at each intrafield site, creating a "fingerprint" of each lens. This method provides greater accuracy in determining the optical response of an exposure tool than field-by-field analysis and can be used to calculate interfield and intrafield precision[6]. LEMSYS™ uses the base set of lens fingerprints and vector map subtraction to calculate stepper to stepper lens fingerprints. Calculating intrafield mismatches explicitly from differential lens analysis requires that (N*(N-1)/2) lens fingerprints be generated for N steppers. By using this inference technique LEMSYS™ was able to calculate the 231 lens fingerprints needed by performing vector map subtraction for every possible pairing of steppers.

3. GRID RESULTS

A summary of the grid model coefficients for the ASML 5500/60 steppers is listed in table 1 and plotted in figure 3. Of concern were the translation and orthogonality terms for ASML stepper 01 and stepper 02. These errors were verified by running the standard preventive maintenance metrology procedure using artifact wafers. The coefficients derived in MONO-LITH® were then entered into the stepper configurations to offset these distortions. Plots of the resulting stepping grids are displayed in figure 4 and the new factors listed in table 2. Table 1 also displays high orthogonality values on both ASML 2500/40s. However no attempt was made to correct these errors as mixing of products between the 5500/60s and 2500/40s is not done, the field-by-field alignment of the Ultratech's minimized their impact, and correction would have a negative effect on the product already in-line.

The grid model results of the Ultratech steppers are also listed in table 2. The translation vector errors for these tools range from -258.8nm to 325nm and, as stated in the previous section, indicate their alignment precision. Although these values do not exceed our non-critical overlay specification they are still quite large. As a result we are currently modifying the 1X alignment schemes and reviewing each system's configuration.

4. LENS MATCHING RESULTS

The results of LEMSYS™ 1X to 5X lens matching modeling are shown in tables 3 and 4; listing thirteen matched pairs with errors worse than 0.3μm. The largest overlay errors were predominantly between the 1X steppers ULT914 and ULT910 and the ASML systems. Stepper ULT9144 has had a history of lens problems and has since been shut down. The lens fingerprint and distortion analysis of ULT910 indicated a large magnification error. During investigation it was discovered that the nitrogen gas to the reticle cooling bar had been disconnected. The thermal expansion of the evaluation reticle during exposure generated the large magnification errors. Removing ULT9144 and ULT910 from the population, lens mismatches are more randomly distributed among the steppers and the number of matched pairs with overlay errors greater than 0.3μm is reduced to three.

5. EXPANSION TO WIDEFIELD AND NON-CONCENTRIC MATCHING

The 1X Ultratech widefield reticle layout provides for three individual fields per reticle. Since field 1 was reserved for the 14mm X 14mm die used in concentric field matching we used fields 2 and 3 to included alignment mark arrays for widefield and non-concentric field characterization. A 27mm X 14mm array of 155 ASML alignment marks, figure 5, was drawn in field 2. When imaged, this array maps the usable lens area of the Ultratech widefield stepper. This size of exposure field is much larger than that of the ASML steppers so a 5X metrology job was created segmenting this exposure area into two 13.5mm X 14mm fields. The stage registration data was then recorded and spreadsheeted in the same manner as the concentric matching process. The full lens fingerprint of ULT914 is plotted in figure 6.

The final reticle field contained an 18 X 7 array of alignment marks that will be used to evaluate the maximum 27mm X 9mm rectangular area of the widefield lens, figure 1. As the purpose of this array is to characterize non-concentric field overlay it was specifically designed to be segregated into two 13.5mm X 9mm die, each a 9 X 7 array of alignment marks. The alignment marks of two 5X exposure fields will exactly overlay the marks of the 1X field, figure 7. In non-concentric field matching some interfield and intrafield parameters, such as die magnification, die rotation, and grid scaling, are not accurately characterized using conventional lens and stepper models. Referencing the 1X and 5X registration data sets to the ASML stage coordinate system each field can first be modeled independently in MONO-LITH®. Software optimized data sets can then be generated for each field and overlaid for non-concentric matching. This approach will allow use of classical models for overlay characterization until algorithms for non-concentric matching are fully developed.

6. CONCLUSION

This paper details a procedure capable of measuring, modeling, and estimating improvements of concentric imaging fields in a mix-and-match stepper population that can be effectively used to minimize total overlay errors. The approach we have taken uses the high precision stage of the ASML 5500/60 stepper to establish an absolute coordinate system. Intrafield and interfield distortion data from each stepper was referenced to this coordinate system using reticles that imaged a 9 X 9 array of ASML alignment marks. The data was then modeled for grid coefficients using the MONO-LITH® registration software package. LEMSYS™, an extension to MONO-LITH®, was next used to generate maximum lens overlay error values for each pair of steppers. Although the goal of this study was to characterize the overlay tolerances of our current mix-and-match process some tuning was performed using the error factors derived by MONO-LITH®. Expansion of this technique to 1X full field and non-concentric field matching was also discussed. Overall, we have determined that our existing equipment set is capable of running within our present overlay specification, however, further improvements are being pursued.

ACKNOWLEDGEMENTS

The authors would like to thank Ursula Caldwell, Barbara Gerber, and Dixie Hill of AMI's ALFA area and Scott Donaldson, of the photolithography group, for their work in exposing and measuring the test wafers. We would also like to thank James Wacker, formerly of ASML, for his assistance in accessing and formatting the stage registration data. Thanks also to Ed Orr, Kirk Brown, Glenn Davis, and Kajsa Norgren for reviewing and commenting on this paper.

REFERENCES

1. A. Charles, F. Sundermann, L. Garcia, and D. Djaber, "Wide field stepper in a mix and match production environment", *Proceedings of the OLIN microlithography seminar*, pp. 205-222, 1996.
2. W. Flack, G. Flores, J. Pellegrini, and M. Merrill, "An optimized registration model for 2:1 stepper field matching", *SPIE proceedings volume 2197*, pp. 773-752, 1994.
3. M. Perkins and Jonathan Stamp, "Intermix technology: the key to optimal stepper productivity and cost efficiency", *Microlithography World*, 1993.
4. M. van der Brink, C. deMol, and R. George, "New 0.54 aperture i-line wafer stepper with field by field leveling combined with global alignment", *SPIE proceedings volume 921*, pp. 180-197, 1988
5. S. Wittenkoek, J. van der Werf, and R. George, "Phase gratings as wafer stepper alignment marks for all process layers", *SPIE symposium on optical microlithography IV*, 1985
6. T. Zavecz, "Characterization and tuning", *SPIE symposium on microlithography short course 13*, 1996.

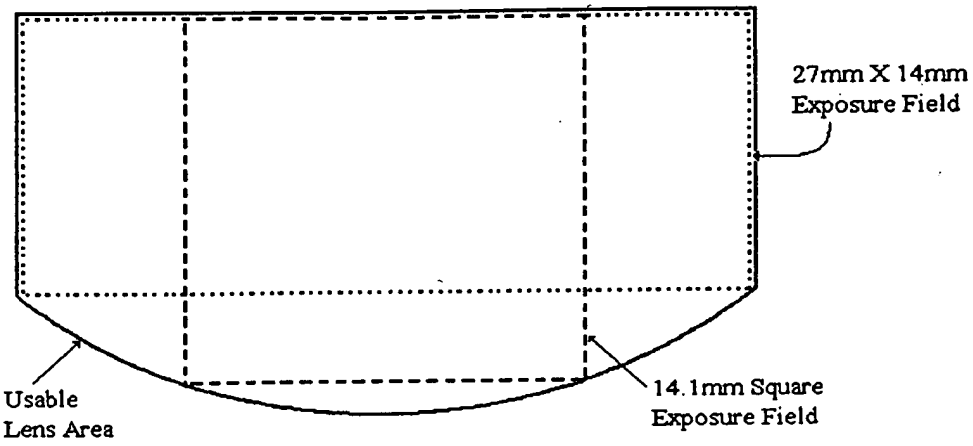


Figure 1: Ultratech's usable lens area and maximum square and rectangular exposure areas.

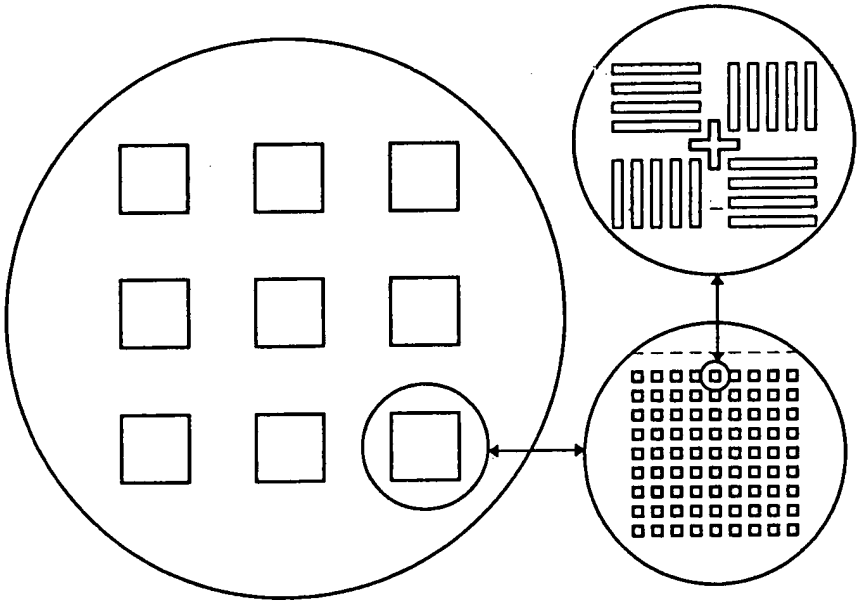


Figure 2: Grid stepping pattern and alignment mark array field.

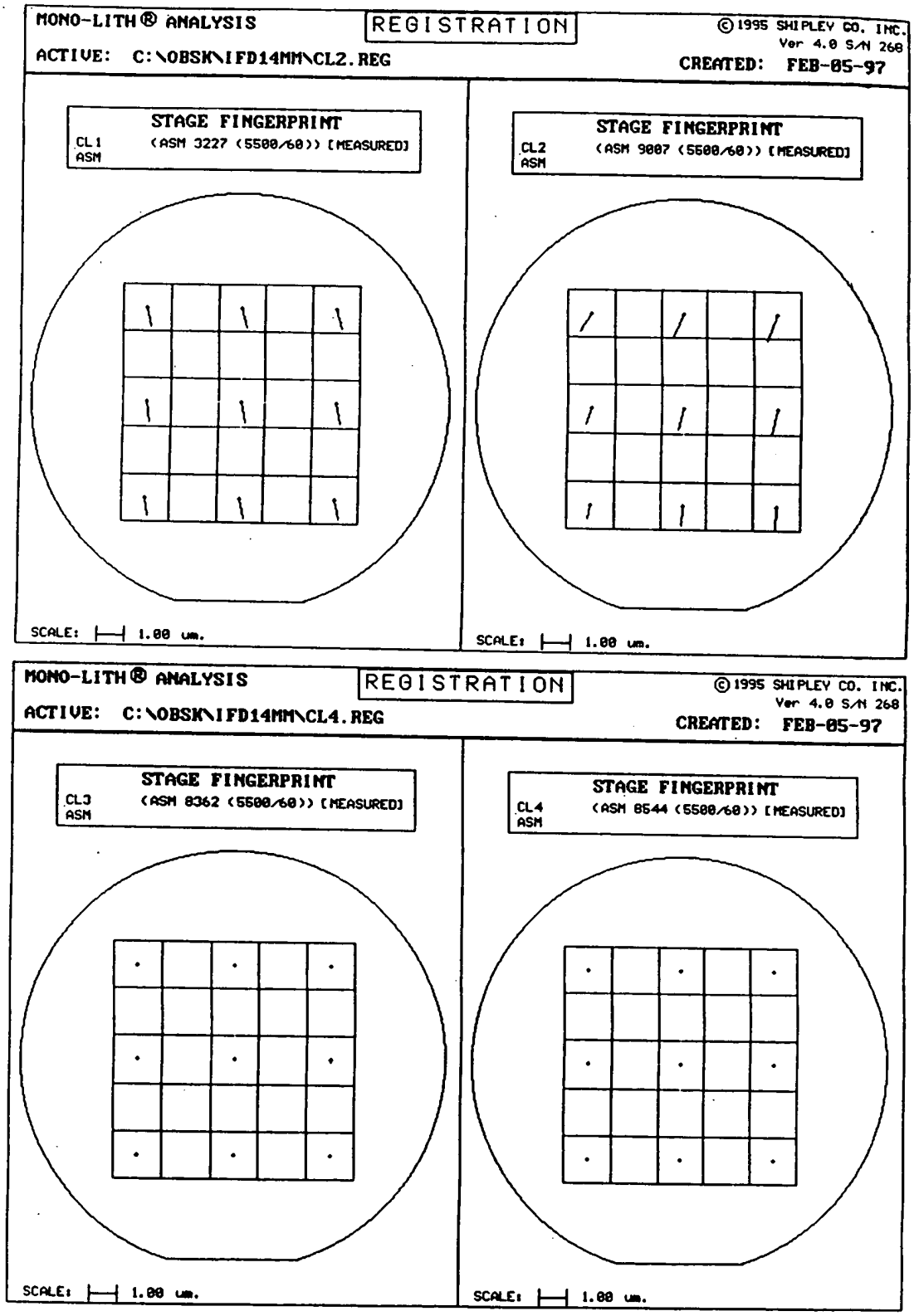


Figure 3: Uncorrected grid plots of ASML 5500/60 steppers, ASM 01, 02, 03, 04.

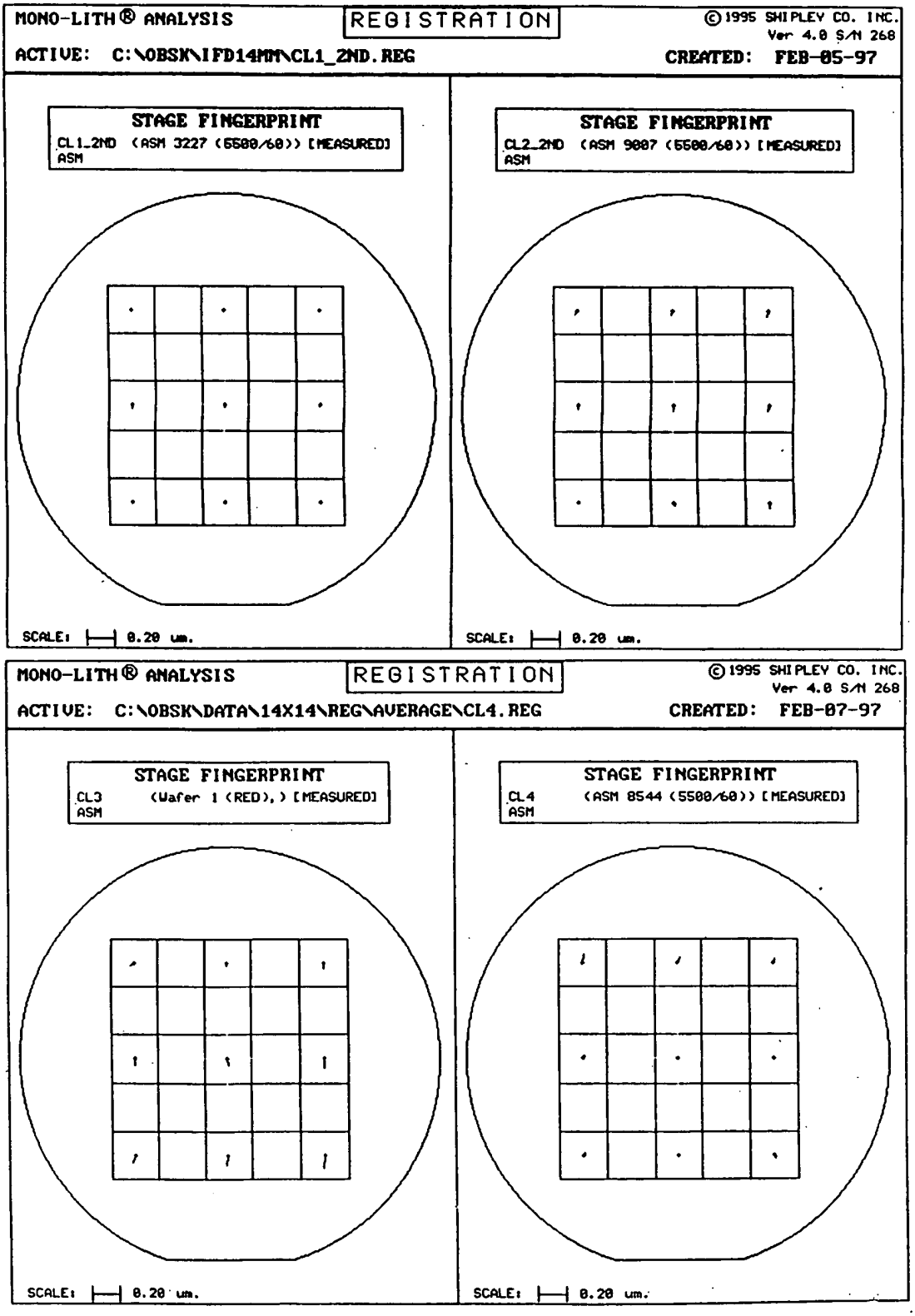


Figure 4: Optimized grid plots of ASML 5500/60 steppers, ASM 01, 02, 03, 04.

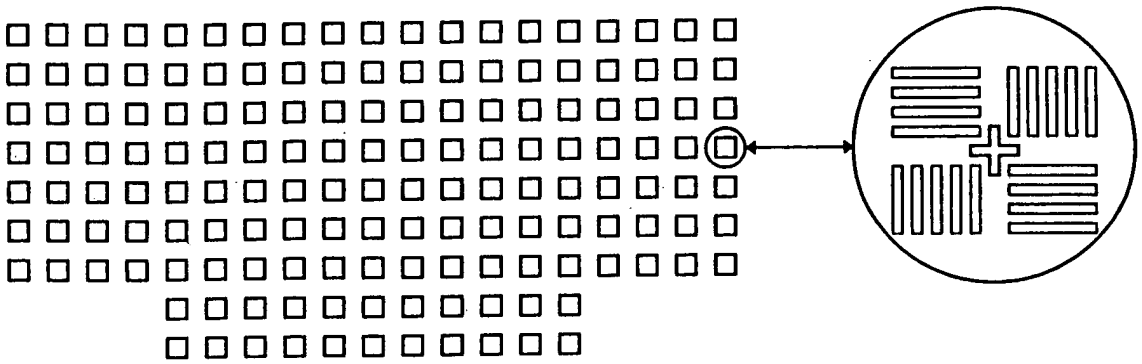


Figure 5: Ultratech widefield matching array. 155 total marks spaced 1.5mm apart on reticle.

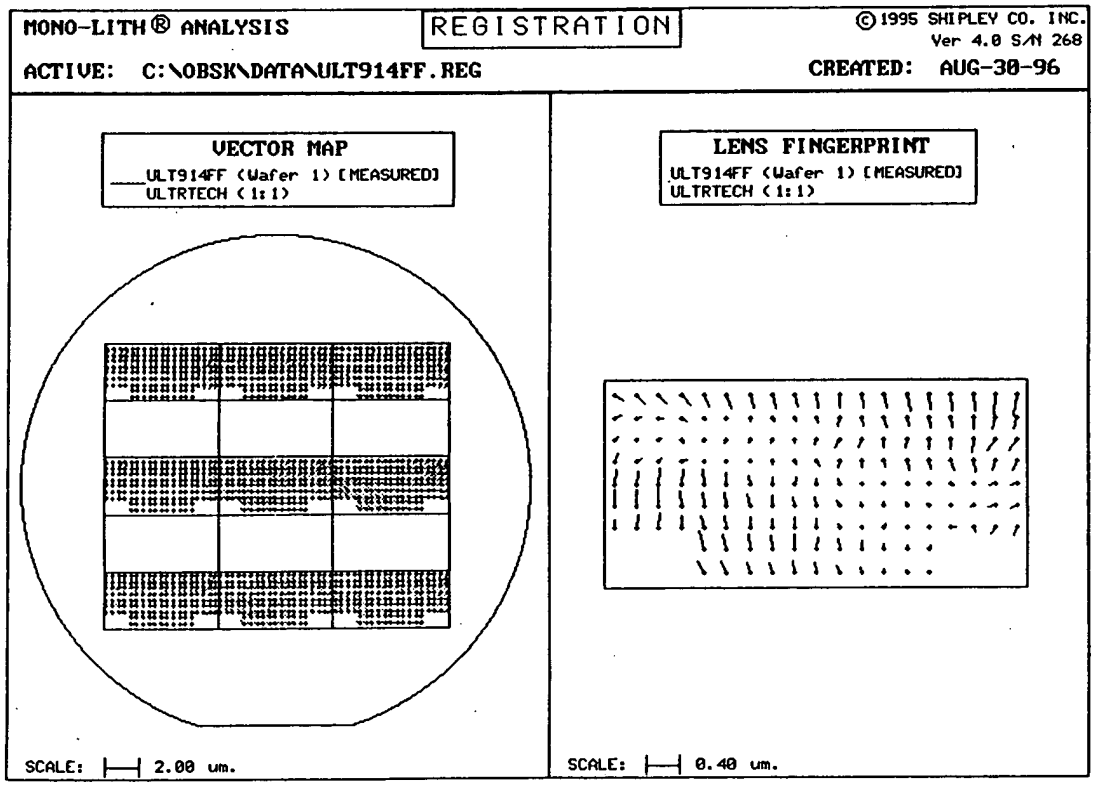


Figure 6: Vector plot and lens fingerprint of 1X widefield.

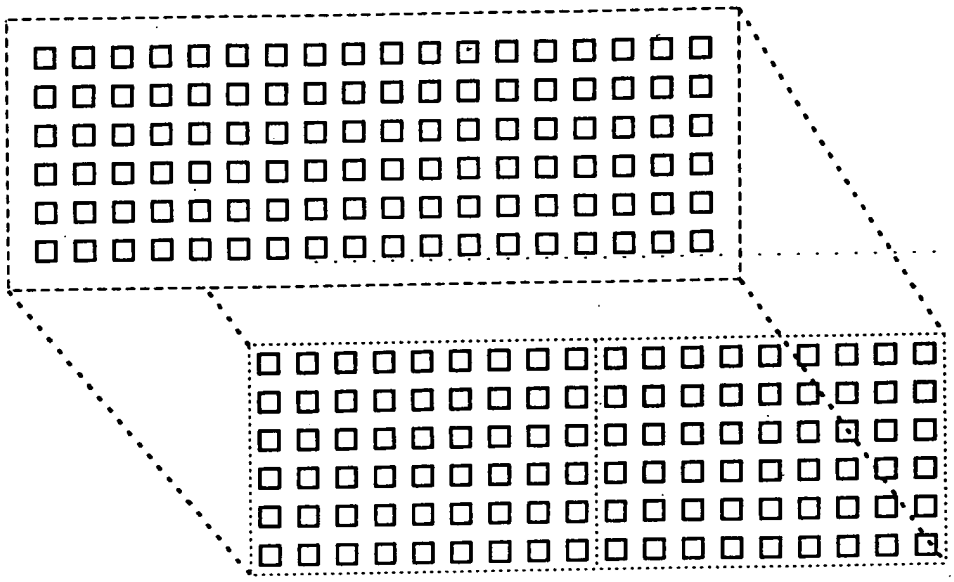


Figure 7: Single 1X field overlaying two 5X fields.

	Translation (nm)		Rotation	Scale (nm/mm)		Ortho.	Residual (nm)	
	X	Y	(nm/mm)	X	Y	(nm/mm)	X	Y
ASM01	145.5	-696.5	-0.147	0.647	0.196	-0.263	36	20
ASM02	-207	-736.8	-5.503	0.541	-1.421	10.612	36	23
ASM04	-0.5	-41.8	-0.49	0.067	0.081	0.459	24	20
ASM03	8.5	12.6	-0.544	-0.045	0.09	0.556	37	32
ASM21	-15.5	-10.1	-0.4	-0.096	-0.049	-9.07	25	30
ASM22	62.7	-40.1	0.174	0.048	-0.081	-9.521	40	35

Table 1: As measured modeled grid vector errors for 5X steppers.

	Translation (nm)		Rotation	Scale (nm/mm)		Ortho.	Residual (nm)	
	X	Y	(nm/mm)	X	Y	(nm/mm)	X	Y
ASM01	-4.5	-6.5	0.003	-0.003	-0.004	-0.003	35	17
ASM02	-8	-26.3	-0.353	0.046	-0.011	0.562	23	72
ASM04	-0.5	-41.8	-0.49	0.067	0.081	0.459	24	20
ASM03	8.5	12.6	-0.544	-0.045	0.09	0.556	37	32
ASM21	-15.5	-10.1	-0.4	-0.096	-0.049	-9.07	25	30
ASM22	62.7	-40.1	0.174	0.048	-0.081	-9.521	40	35
ULT901	-233.5	-38.2	1.294	0.505	1.216	-2.3	63	54
ULT902	113.6	-132.8	1.492	-0.635	1.678	-2.625	47	52
ULT903	228.8	-191	2.263	-0.158	1.746	-3.011	57	76
ULT904	-171.7	-104.8	0.571	-1.305	1.937	-0.857	61	59
ULT905	47.6	6	1.217	0.852	0.432	-2.924	48	52
ULT907	240.7	159.1	-0.136	0.678	2.033	-0.429	70	57
ULT909	-68	-17.1	1.229	-0.804	2.092	-1.781	77	65
ULT910	111.5	-52.6	1.595	-0.514	1.487	-2.386	68	75
ULT913	151.1	-258.8	0.675	-1.924	0.96	-1.951	47	82
ULT914	-196.6	203.9	0.813	-0.4461	1.115	-1.059	51	57
ULT915	50.7	-19.8	-0.057	-0.666	2.278	-1.448	51	62
ULT916	69	-39.3	1.195	-1.026	0.6	-1.579	61	66
ULT931	-151.4	-192.7	0.649	0.483	1.565	-1.467	36	84
ULT932	166.7	325	-0.222	-0.45	1.186	0.544	48	52
ULT933	255.9	-81.1	-0.63	-0.607	0.955	0.582	54	53
ULT934	110.9	-69.1	0.065	-0.805	0.568	-0.934	46	45

Table 2: As measured modeled grid vector errors for 5X and 1X steppers after correcting grid errors on ASM steppers 01 and 02.

	ASM01	ASM02	ASM03	ASM04	ASM21	ASM22
ASM01		(88 / 82)	(84 / -45)	(30 / 64)	(118 / 104)	(230 / -195)
ASM02	(-88 / -82)		(-61 / 92)	(-93 / -106)	(-109 / 99)	(-178 / -252)
ASM03	(-84 / 45)	(61 / -92)		(-77 / 64)	(-71 / 96)	(152 / -183)
ASM04	(-30 / -64)	(93 / 106)	(77 / -64)		(111 / 143)	(226 / -240)
ASM21	(-118 / -104)	(109 / -99)	(71 / -96)	(-111 / -143)		(154 / -156)
ASM22	(-230 / 195)	(178 / 252)	(-152 / 183)	(-226 / 240)	(-154 / 156)	
ULT931	(-150 / -125)	(-91 / -133)	(-68 / -140)	(-143 / 134)	(-65 / 144)	(161 / -243)
ULT932	(-226 / 253)	(-168 / 258)	(-142 / 276)	(-220 / 317)	(-133 / 198)	(103 / -176)
ULT933	(189 / 149)	(254 / 203)	(-198 / 167)	(-184 / 135)	(-220 / 215)	(267 / 260)
ULT934	(-147 / -138)	(178 / 162)	(153 / -154)	(-123 / -114)	(-196 / 167)	(271 / -253)
ULT901	(-121 / -139)	(106 / -171)	(118 / -131)	(-126 / -178)	(-107 / 112)	(249 / -182)
ULT902	(218 / 115)	(229 / 160)	(222 / 102)	(248 / 159)	(-222 / 120)	(140 / -130)
ULT903	(223 / 136)	(289 / 176)	(233 / 122)	(-225 / 159)	(-238 / 167)	(224 / 203)
ULT904	(-160 / -142)	(-153 / 138)	(140 / 138)	(166 / 181)	(-155 / -105)	(136 / -123)
ULT905	(-166 / -270)	(132 / 271)	(87 / 288)	(-159 / 330)	(-124 / 212)	(127 / -195)
ULT907	(245 / 169)	(311 / 232)	(252 / 185)	(232 / 158)	(-214 / 229)	(284 / 265)
ULT909	(-189 / -200)	(224 / -217)	(165 / -188)	(191 / -213)	(-210 / -166)	(-116 / 132)
ULT910	(286 / 296)	(352 / 359)	(293 / 311)	(-288 / 284)	(-302 / 362)	(354 / 392)
ULT913	(-110 / -105)	(170 / 157)	(110 / 121)	(123 / 108)	(-128 / 162)	(178 / 215)
ULT914	(-283 / 312)	(-310 / 272)	(-276 / 303)	(-253 / 348)	(-333 / 272)	(-396 / 229)
ULT915	(-116 / 90)	(151 / 138)	(125 / 100)	(-85 / 142)	(-170 / -69)	(250 / -136)
ULT916	(-119 / -108)	(152 / -116)	(-111 / -124)	(-111 / 92)	(-140 / 118)	(185 / -177)

Table 3: Maximum overlay vectors from all match pairs. Table units = nm.

BEST MATCHES			WORST MATCHES		
ULT913	ASM01	(-110 / -105)	ULT914	ASM22	(-396 / 229)
ULT916	ASM04	(-111 / 92)	ULT910	ASM22	(354 / 392)
ULT901	ASM21	(-107 / 112)	ULT910	ASM21	(-302 / 362)
ULT915	ASM01	(-116 / 90)	ULT910	ASM02	(352 / 359)
ULT916	ASM01	(-119 / -108)	ULT914	ASM04	(-253 / 348)
ULT913	ASM03	(110 / 121)	ULT914	ASM21	(-333 / 272)
ULT934	ASM04	(-123 / -114)	ULT905	ASM04	(-159 / 330)
ULT913	ASM04	(123 / 108)	ULT932	ASM04	(-220 / 317)
ULT916	ASM03	(-111 / -124)	ULT914	ASM01	(-283 / 312)
ULT915	ASM03	(125 / 100)	ULT910	ASM03	(293 / 311)
ULT901	ASM03	(118 / -131)	ULT907	ASM02	(311 / 232)
ULT909	ASM22	(-116 / 132)	ULT914	ASM02	(-310 / 272)
ULT931	ASM02	(-91 / -133)	ULT914	ASM03	(-276 / 303)
ULT904	ASM 22	(136 / -123)	ULT910	ASM01	(286 / 296)
ULT901	ASM01	(-121 / -139)	ULT903	ASM02	(289 / 176)
ULT902	ASM22	(140 / -130)	ULT910	ASM04	(-288 / 284)
ULT904	ASM03	(140 / 138)	ULT905	ASM03	(87 / 288)
ULT931	ASM03	(-68 / -140)	ULT907	ASM22	(284 / 265)
ULT916	ASM21	(-140 / 118)	ULT932	ASM03	(-142 / 276)
ULT915	ASM04	(-85 / 142)	ULT905	ASM02	(132 / 271)
ULT931	ASM04	(-143 / 134)	ULT934	ASM22	(271 / -253)
ULT931	ASM 21	(-65 / 144)	ULT905	ASM01	(-166 / -270)

Table 4: Best and worst matched pairs maximum overlay vectors. Table units = nm.

Sampling strategy and model to measure and compensate the overlay errors

Chen-Fu Chien [a] , Kuo-Hao Chang [a] , Chih-Ping Chen [b]
[a] Department of Industrial Engineering and Engineering Management, NTHU
[b] Macronix International Co., Ltd., Taiwan, R.O.C.

ABSTRACT

Overlay is one of the key designed rules for producing VLSI devices. In order to have a better resolution and alignment accuracy in lithography process, it is important to model the overlay errors and then to compensate them into tolerances. This study aimed to develop a new model that bridges the gap between the existing theoretical models and the data obtained in real settings and to discuss the overlay sampling strategies with empirical data in a wafer fab. In addition, we used simulation to examine the relations between the various factors and the caused overlay errors. This paper concluded with discussions on further research.

Key words: overlay, stepper, yield improvement, decision analysis, semiconductor manufacturing

1. INTRODUCTION

In semiconductor manufacturing, microlithography is one key technique involved in wafer fabrication processes. Recently, wafer step-and-repeat systems have replaced scanning projection as the photolithographic exposure tool for the fabrication of VLSI (Very Large Scale Integration) devices. In order to function properly, the patterned layers in the fabrication of VLSI devices must overlay each other to within the tolerance that is incorporated in the IC design. With the use of advanced step-and-repeat projection aligners, it is required to increase fine resolution and alignment accuracy. In order to improve the yield, it is important to increase the performance of the stepper and to control overlay errors within the tolerance so as to have a better resolution and alignment accuracy in lithography process. Overlay error is defined to be the displacement error of an exposed photo image field relative to a previously exposed image field [7].

To diagnose the causes of overlay errors and to control the errors through compensation to remove correctable errors are important problems in semiconductor manufacturing. The challenges are even greater with the 0.18-micron technology. There have been a number of related studies on the factors causing the overlay errors [13], the mathematical models [4], and the overlay error control methods [9]. However, few existing models have been applied in practice. Firstly, the variables considered in the existing overlay models are different from the real data measured through the calibration system. There is a need to bridge the gap between theoretical models and the data in real settings. Secondly, because of

Metrology, Inspection, and Process Control for Microlithography XV, Neal T. Sullivan, Editor,
Proceedings of SPIE Vol. 4344 (2001) © 2001 SPIE · 0277-786X/01/$15.00

the operating costs for measuring overlays, the number of overlays measured in one wafer is restricted. In order to obtain the higher overlay accuracy within the limited number of measured overlays, there is a need to develop systematic sampling strategies in which one can decide the number and the corresponding locations of overlays to measure.

This study aimed to review the existing models and to develop a new model that bridges the gap between theoretical models and the data in real settings. We also designed an experiment and got empirical data from a wafer fab to compare the various sampling strategies with the existing strategy for measuring overlay errors. In addition, we used simulation to examine the relations between overlay errors and various factors for validation.

2. LITERATURE REVIEW

2.1 Modeling the overlay error

With the increase of VLSI feature size, the factors causing overlay errors become more complex. According to Schmidt [13], the factors that cause overlay errors belong to the system, stepper, reticle accuracy, linewidth variation, and wafer distortion (see Table1).

The following terminology and notations are generally used in the study.

(x, y) : Intrafield coordinate system, With respect to the center of a field

(X, Y) : Interfield coordinate system, with respect to the center of the wafer

d_x, d_y : The intrafield overlay errors in the field coordinate system, (x, y).

d_X, d_Y : The intrafield overlay errors in the wafer coordinate system, (X, Y).

d_{x+X} : The sum of the intrafield and interfield overlay errors

d_{y+Y} : The sum of the intrafield and interfield overlay errors

T_x, T_y : The intrafield translation

T_X, T_Y : The interfield translation

T_{x+X} : The sum of the intrafield

T_{y+Y} : And interfield translation

E_X, E_Y : The interfield expansion

M_x, M_y : The intrafield magnification

R_x, R_y : The intrafield rotation

R_X, R_Y : The interfield rotation

r_x, r_y : The intrafield residual

r_X, r_Y : The interfield residual

Brink et al. [4] developed an overlay model that considered both the interfield and intrafield effects. The intrafield model was based on MacMillen and Ryden [8] and added the fifth-order lens distortion variables as follows:

$$d_x = T_x + M_x x - R_x y - T_{xx}x^2 - T_{yx}xy + W_x y^2 + D_{3x}x(x^2+y^2) + D_{5x}x(x^2+y^2)^2 + r_x \quad (1)$$

$$d_y = T_y + M_y x - R_y y - T_{xy}x^2 - T_{yy}xy + W_y x^2 + D_{3y}y(x^2+y^2) + D_{5y}y(x^2+y^2)^2 + r_y \quad (2)$$

where

T_{xx}, T_{xy} : The tilt coefficients of the mask
T_{yx}, T_{yy} : The tilt coefficients of the mask
W_x : The wedge distortion in X-axis direction
W_y : The wedge distortion in Y-axis direction
D_{3x}, D_{5x} : The third order distortion coefficients
D_{3y}, D_{5y} : The fifth order distortion coefficients

Table 1. Overlay Error Factors

Causes	Overlay errors
System	Vibration Temperature
Stepper	Alignment error Lens distortion (optical stepper only) Wafer clamping Wafer table errors
Reticle accuracy	Reticle in-plane distortion Pattern placement errors Reticle clamping
Linewidth variation	Wafer (exposure, development, etching, etc.) Reticle (exposure, development, etching, etc.)
Wafer distortion	Flatness/curvature Pattern movement/slip

Table 2. The existing overlay models

Author (year)	Parameters in the model
Perloff (1978)	Translation Rotation Expansion
MacMillen and Ryden (1982)	Translation Rotation Expansion lens trapezoid lens distortion
Peski (1982)	scale factors array orthogonality lens distortion magnification Ratio
Arnold (1983)	Translation Rotation Expansion lens trapezoid lens distortion bow coefficients
Brink et al. (1988)	Translation Rotation Expansion lens trapezoid three-order lens distortion fifth-order lens distortion bow coefficients
Lin and Wu (1998)	Translation Rotation Magnification mask tilt wedge distortion high-order distortion
Chien et. al (2000)	Translation rotation exoansion non-orthogonality

The interfield model combined the model developed by Perloff [10] and the bow parameters considered in Arnold [3]

as follows:

$$d_X = T_X + E_X X - R_X Y + B_X Y^2 + r_X \quad (3)$$

$$d_Y = T_Y + E_Y Y + R_Y X + B_Y X^2 + r_Y \quad (4)$$

Lin and Wu combined the intrafield and interfield errors in x, y directions, respectively.

$$\begin{aligned} d_{x+X} = {} & T_X - R_X Y + M_X X + B_X Y^2 \\ & + T_x - R_x + M_x x - T_{xx} x^2 - T_{yx} xy + W_x y^2 \\ & + D_{3x} x(x^2 + y^2) + D_{5x} x(x^2 + y^2)^2 \end{aligned} \quad (5)$$

$$\begin{aligned} d_{y+Y} = {} & T_Y + R_Y X + M_Y Y + B_Y X^2 \\ & + T_y + R_y x + M_y y - T_{xy} xy \_ T_{yy} y^2 + W_y x^2 \\ & + D_{3y} y(x^2 + y^2) + D_{5y} y(x^2 + y^2)^2 \end{aligned} \quad (6)$$

where B_x, B_y denote the stage bow coefficients.

We proposed the interfield overlay error as the following:

$$d_X = T_X + E_X X - (N + \theta)Y + r_X \quad (7)$$

$$d_Y = T_Y + E_Y Y - (\theta - N)X + r_Y \quad (8)$$

In particular,

$$N = \frac{R_X - R_Y}{2} \quad (9)$$

$$\theta = \frac{R_X + R_Y}{2} \quad (10)$$

Similarly, to match the real setting, they proposed the intrafield overlay models as the following:

$$d_x = T_x + M_x' x - R_x y + r_x \quad (11)$$

$$d_y = T_y + M_y' y + R_y x + r_y \quad (12)$$

where the variable, M_x', is defined by $M_x' = \Delta M = M * M_x$ when the designed lens magnification M is given. In other words, M_x' is multiple of M_x and the multiple depends on the designed lens magnification.

The interfield and intrafield overlay models are combined as follows:

$$d_{x+X} = T_{x+X} + E_X X - (N + \theta)Y + M_x' x - R_y y + r_{x+X} \quad (13)$$

$$d_{y+Y} = T_{y+Y} + E_Y Y - (\theta - N)X + M_y' y + R_y x + r_{y+Y} \quad (14)$$

The proposed model has the advantage that bridges the empirically assessable data and the variables in real settings. It can be empirically derived through statistical analysis of assessed data.

Table 2 summarizes the existing overlay models. However, to consider all the factors is very difficult and costly in wafer fabs. Thus, one feasible way is to start with focusing on the critical and correctable factors that the steppers can measure and compensate.

3. SAMPLING STRATEGIES

3.1 Framework

The sampling strategies involve two phases. The first is to determine the corresponding intrafield overlay locations. The second is to consider the interfield overlay locations. Then we consider total number of sampled overlays in a wafer and its corresponding R-square. In order to determine the corresponding sampling locations, we design eight kinds of sampling

locations in a field to compare goodness of fit to the overlay model. Furthermore, we examine the relation between the numbers of sampled overlays and the corresponding R-square that is the highest among different sampling locations. We pointed out the relation between the sampling number and the corresponding R-square in a figure to illustrate the tradeoffs between sampling cost and compensation effectiveness.

3.2 A technical insight

From equation 13 and 14, overlay error is dependent with the intrafield coordinates (x, y). We transform it to polar coordinates (r, θ).

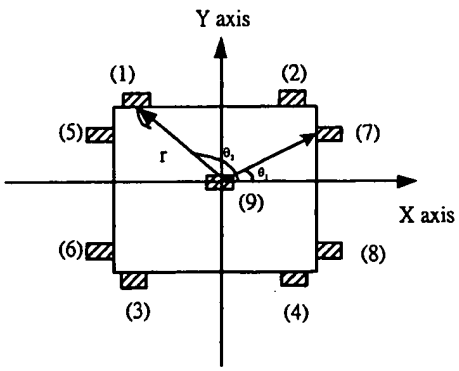


Figure 1 The intrafield sampling

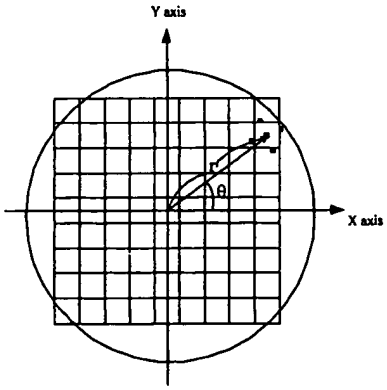


Figure 2 The interfield sampling

As Figure 1, the intrafield radius is defined to the distance of overlay and central overlay. The angle is defined as Figure 1. Observing Figure 1 overlay (1), (2), (3) and (4) represent the information of X direction overlay accuracy. Similarly, overlay (5), (6), (7), and (8) represent the information of X direction overlay accuracy. From intrafield equation (11) and (12), parameter R_x is related to y and parameter R_y is related to x. Therefore, X direction and Y direction are equally important and cannot lose either of them.. Theoretically, we should choose two overlays among (1), (2), (3) and (4). Similarly, two overlays should be chosen among (5), (6), (7), (8). From statistical viewpoints, when the difference of x and y angles $\theta_2 - \theta_1$, is bigger, parameter R_x and R_y should be estimated more precise. It is because when $\theta_2 - \theta_1$ is bigger even one field rotate lightly, they will result in more difference and then the parameters R_x and R_y will be estimated more precise.

In interfield sampling, we also transformed the interfield coordinates (X, Y) into polar coordinates (r, θ). The parameters r and θ are defined as Figures 2. Interfield equations (7) and (8) show that expansion parameters are related to (X, Y) coordinates. In particular, E_X is related to X and E_Y is related to Y. From expansion parameter pattern simulation, we find that outer side overlays will result in bigger overlay errors. Even light expansion happen will result in

bigger overlay errors in the outer side. Therefore, if we choose overlays directly in the outer side, that is, overlay with bigger r, the bigger overlay error can be compensate very well, so does other overlays in the inner side. Due to this reason, we predict choosing overlay in the outer side will have more precise expansion parameters E_X and E_Y. Next empirical study will prove our insight.

4. AN EMPIRICAL STUDY

4.1 Problem structuring

This section presents an empirical study in a wafer fab of a semiconductor company in Taiwan. This company has been one of the Top 500 manufacturers in Taiwan since 1993. Briefly, this fab manufactures a broad line of high-performance non-volatile memory ICs and micro-controller ICs that are used in communication systems, computers, and high-end electronic consumer systems.

Focusing on the real setting in this wafer fab, we developed a new overlay model that used the empirically assessed data as the independent variables. In particular, the steppers can measure and compensate the error variables of translation, expansion, non-orthogonality, and rotation in the fab. Following the idea of experiment of design (DOE), we consider two factors, r and θ. Observing Figure 1, we can classify (1) and (5), (2) and (7), (3) and (6), (4) and (8) into several different classes, respectively. Suppose that totally we have four overlays. Each class should sample one overlay. Theoretically, we have $2 \times 2 \times 2 \times 2 = 16$ combinations. After dismissing the cases that don't contain both X and Y information, the designed intrafield overlay sampling are illustrated as Figure 3. For demonstration, we conducted experiments to obtain real data in the fab. We used the stepper compensating function to create four different settings on the experiment wafers that involved (1) the both interfield and intrafield effects, (2) the interfield effects only, (3) the intrafield effects only, and (4) no effect, respectively.

4.2 Intrafield sampling strategies

We compared the R-square of these eight designed intrafield sampling. Totally we sampled 51 fields in a wafer. Actually, the existing sampling location in the fab is illustrated in Figure 3 (1).

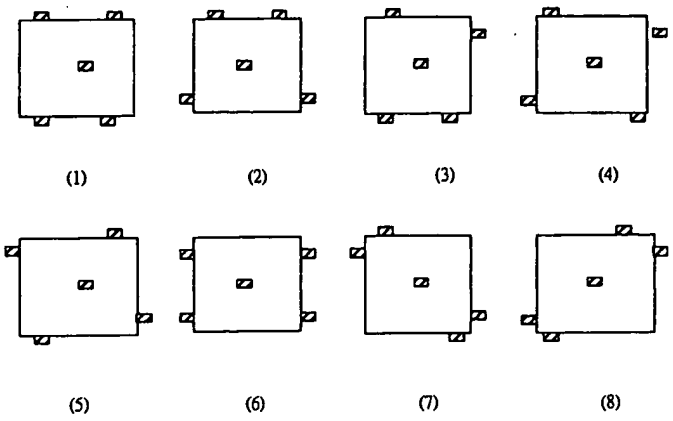


Figure 3 The intrafield sampling locations

By applying multiple linear regression analysis [5] to fit the proposed models as given in equations 13 and 14., we estimate the parameters and their R-squares.

4.3 Interfield sampling strategies

To evaluate various sampling strategies, we designed an interfield sampling experiment by considering two factors of radius and angle. This experiment included nine wafers with nine different interfield sampling strategies (see Figure 4-12). According to the simulation results, the field with bigger radius will result in bigger overlay errors no matter rotation or expansion parameters. Therefore, we designed a wafer with two different patterns which sampling overlays mainly be chosen in the outer fields. (see Figure 4 and Figure 6). Secondly, we compared the overlay error of different radius (see Figure 4, Figure 5 and Figure 6 and Figure 7). As Figure 4 and 5, these two sampling strategies have the same angle but different radius. Alternatively, Figures 6 and 7 have the same radius but different angles. Thirdly, we combined some parts of sampling pattern (1) (Figure 4) with some parts sampling pattern (3) (Figure 6) to create sampling type (8) (Figure 11). Another different interfield sampling is designed as Figure 12. In the empirical study given in the next chapter, we will compare the effect of different interfield sampling strategies and their goodness of fit to the overlay models.

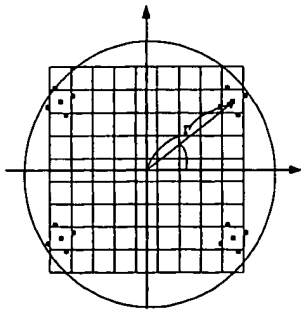


Figure 4 The interfield sampling pattern (1)

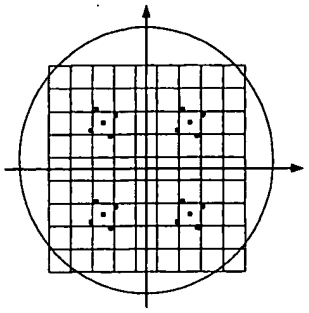


Figure 5 The interfield sampling pattern (2)

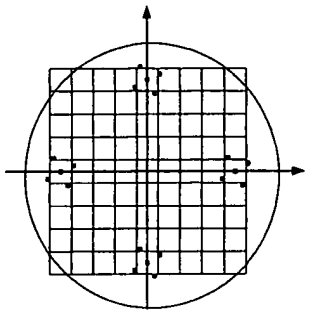


Figure 6 The interfield sampling pattern (3)

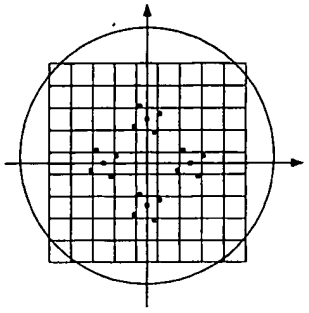


Figure 7 The interfield sampling pattern (4)

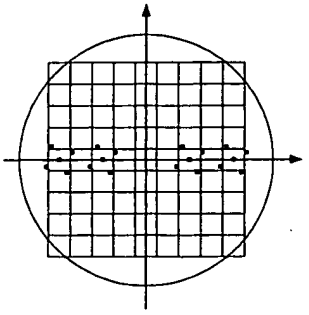


Figure 8 The interfield sampling pattern (5)

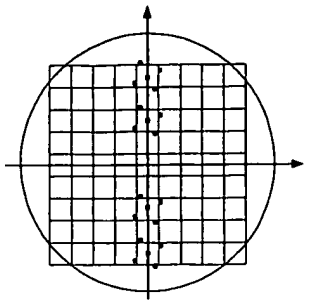


Figure 9 The interfield sampling pattern (6)

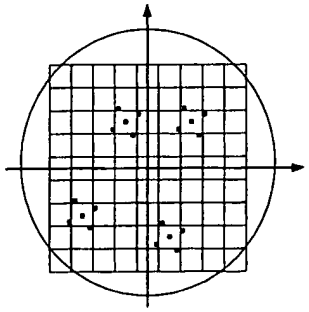


Figure 10 The interfield sampling pattern (7)

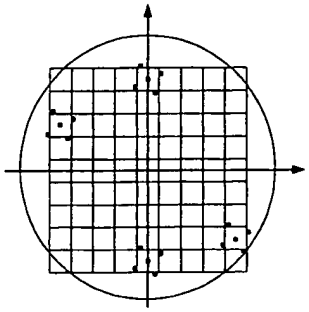


Figure 11 The interfield sampling pattern (8)

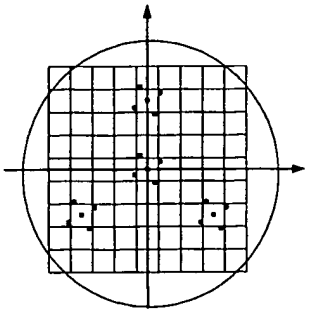


Figure 12 The interfield sampling pattern (9)

4.4 Results

From this empirical study, we found the intrafield sampling type (4) is better than others are (see Table 3). Theoretically, the parameter R_x and R_y will be estimated more precise if the $\theta_2 - \theta_1$ is bigger and the estimated parameters M'_x and M'_y will be closer to the real parameters if the radius r is bigger. The intrafield sampling type (4) has both advantages. On one hand, it has bigger $\theta_2 - \theta_1$ and bigger radius. On the other hand, its sampling pattern possesses symmetric characteristics. The empirical study confirmed this result.

Table 3. R-square of different sampling location and various number of sampling overlays

# of sampling \ Sampling Type	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
255	99.4%	99.7%	99.4%	99.6%	99.8%	99.7%	99.6%	99.8%
100	97.2%	94.7%	96.1%	98.1%	99.3%	95.2%	93.1%	95.4%
50	95.8%	97.2%	96.7%	98.3%	96.5%	95.1%	90.6%	94.1%
20	87.1%	96.8%	92.1%	96.5%	91.0%	88.5%	76.7%	81.1%
10	71.5%	83.1%	61.2%	83.6%	83.1%	73.8%	52.5%	69.7%

As shown in Table 3, we found that there is no significant difference among the different sampling locations as the number of sampled overlays was large (255 and 100). However, there were differences among the different sampling locations as the number of sampled overlays was relatively small (i.e., 10 and 20). Considering the existing sampling number (i.e., 20 overlays), the sampling location shown in Figure 3 (4) had higher goodness of fit than the existing sampling location. The residual analysis of the sampling location (4) (see Figure 13) validated the model.

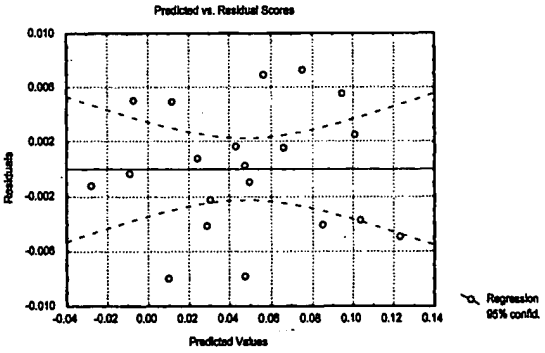


Figure 13. Residual analysis

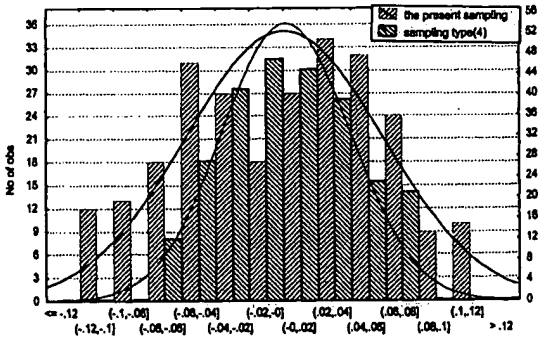


Figure 14. The histogram of residuals

Indeed, the validity of sampling strategy depends on the overlay error after compensation. The R-square denotes the proportion of total variability in the response variable that is explained by the independent variables. That is, R-square denotes the goodness of fit of the model. For validation, we used the estimated parameters of the proposed overlay models to compensate the overlay errors and compared the residuals between the present sampling location shown in Figure 3 (1) and the better sampling location shown in Figure 3 (4). Figure 14 summaries the results in histogram. As shown in Figure 14, the residuals of the sampling location (4) were more concentrated than the existing sampling location. Furthermore, we used the norm of 255 overlay errors after compensation in sampling location (4) and the existing sampling location, i.e. sampling location (1).

Let

$$\| v_i \| = \sqrt{r_1^2 + r_2^2 + \ldots . r_{255}^2} \tag{15}$$

On one hand, the norm of overlay errors after computation that was based on the sampling location (4) is 0.598. On the other hand, the norm of overlay errors after computation that was based on the existing sampling location is 0.924. This implies that the sampling location (4) can eliminate further 35.4% overlay error than the existing sampling location. Furthermore, we examined the relation between the numbers of sampled overlays and the corresponding R-square that is the highest among different sampling locations as shown in Figure 15. Certainly, the more sampled overlays the more information and thus the higher the R-square values. Based on the information given in Figure 15, the decision-maker can tradeoff the sampling number (i.e., the sampling cost) and the R-square that the model possibly fitted (i.e., goodness of fit and compensation).

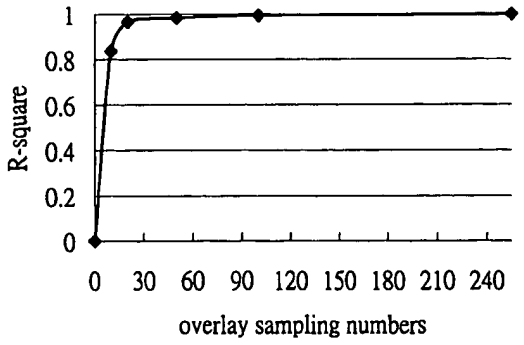


Figure 15. Comparing overlay sampling numbers

From Table 4, In interfield sampling, we have the following findings.

1. When the sampling number is restricted, the bigger radius sampling and symmetric sampling will be better than other sampling.
2. When rotation error exists and its effect is bigger than others apparently, interfield sampling type (1) (see Figures 4 can estimate rotation parameter θ more pecise than other sampling type.
3. When expansion error exists and its effect is bigger than others apparently, interfield sampling type (3) (see Figures 6) can estimate rotation parameter θ more precise than other sampling type.

Thus, the sampling number and the sampling location have a significant impact on the values of estimated parameters of the derived models and on the degrees of compensating the correctable causes. For any overlay model, an inappropriate sampling strategy may cause ill fitness of the model and bad compensation of the errors.

5. CONCLUDING REMARKS

In this study, we reviewed the theoretical overlay models as summarized in Table 2. Based on the existing models and real setting in a fab, we proposed the new intrafield and interfield overlay models that effectively incorporate the assessable data and correctable overlay errors. We designed experiments to obtain data to test different intrafield sampling locations. We found a better sampling location than the existing one that eliminated more 35.4% overlay errors. We pointed out the relation between the sampling number and the corresponding R-squares to illustrate the tradeoffs between sampling cost and compensation effectiveness. Further research is needed to investigate the sampling strategies for interfield locations with regarding the proposed models. Further research is also needed to examine the proposed models and the proposed sampling location in various setting.

ACKNOWLEDGEMENTS

This research is partially supported by Macronix International Co., Ltd. The authors appreciate generous assistance from Wafer Fabrication Business Unit (II). Special thanks go to Dr. Yih-Cheng Shih for his suggestions.

REFERENCES

1. Armitage, J. D. and J. P. Kirk, "Analysis of overlay distortion patterns", *Proceedings SPIE: Integrated Circuit Metrology, Inspection, and Process Control II.* 921, 207-222 ,1988.
2. Arnold, W. H, "Overlay simulator for wafer steppers", *Proceedings SPIE: Optical/Laser Microlithography*, vol. 922, 94-105, 1988.

3. Arnold, W. H., "Image placement differences between 1:1 projection aligners and 10:1 reduction wafer steppers", *Proceedings SPIE: Optical Microlithography*, 394, 87-98, 1983.
4. Brink, M. A., C. G. M. Mol and R. A. George, "Matching performance for multiple wafer steppers using an advanced metrology procedure", *Proceedings SPIE: Integrated Circuit Metrology, Inspection, and Process Control II*, 921, 180-197, 1988.
5. Draper, N. R. and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, 1981.
6. Hasan, T. F., S. U. Katzman and D. S. Perloff, "Automated electrical measurements of registration error in step-and-repeat optical lithography systems ", *IEEE Transactions on Electron Devices*, 27 (12), 2304-2312, 1980.
7. Lin, Z. and W. Wu, "Multiple Linear Regression Analysis of the Overlay Accuracy Model", *IEEE Transaction on Semiconductor Manufacturing*, 12, 229-237, 1999.
8. MacMillen, D. and W.D. Ryden, "Analysis of image field placement deviations of a 5× microlithographic reduction lens", *Proceedings SPIE: Optical Microlithography-Technology*, 334, 78-89, 1982.
9. Magome, N. and H. Kawar, " Stepper stability improvement by a perfect self-calibration system", *Proceedings SPIE* vol.2197, 990-996, 1994.
10. Perloff, D. S. "A four-point electrical measurement technique for characterizing mask superposition errors on semiconductor wafer", *IEEE Journal of Solid-State Circuits*, 13(4), 436-444, 1978.
11. Peski, C. K., "Minimizing pattern registration errors through wafer stepper matching techniques", *Solid State Technology*, 25 (5), 111-115, 1982.
12. Rangarajan, B., M. Templeton, L. Capodieci, R. Subramanian and A. Scranton, "Optimal Sampling Strategies for sub-100nm Overlay", *Proceedings SPIE*, vol.3332, .348-359, 1998.
13. Schmidt, D. and G. Charache, "Wafer process-induced distortion study for x-ray technology", *Journal of Vacuum Science and Technology*, B9 (6), 3237-3240, 1991.

Interfield Sampling Method Dependency of Overlay and Global Alignment

Jinseog Hong, Junghyeon Lee, Hanku Cho, Jootae Moon, and Sangin Lee
(Semiconductor R&D Center, SAMSUNG Electronics Co. Ltd., 449-900 San#24, Nongseo-Ri, Kiheung-Eup, Yongin-Shi, Kyungki-Do, Korea)

ABSTRACT

According to the classical calculation of overlay margin as 1/4 design rule, the overlay control requirement for sub-0.15um design rule device is nominally below 40nm. To meet this demand, it is necessary that one should analyze every part in global alignment and overlay measurement procedure, then factor out the parameter that is known to affect overlay control, and correct it as much as possible. One of the major causes degrading overlay budget seems to be the nonoptimized wafer sampling method. Compensated but undercorrected overlay errors usually fitted as linear terms can be amplified due to improper sampling method e.g. asymmetric one.

In this paper, we have investigated the possible causes that yield global alignment noise and the sampling method dependency of global alignment repeatability and overlay model calculations. The achievement of better alignment repeatability is critical for improving not only in-wafer overlay but wafer-to-wafer overlay control. It is thus evident that overlay control can be improved by reducing alignment noises or by optimizing sampling method. Global alignment repeatability and its results are significantly affected by which chips in a wafer map are selected as global alignment purpose. This result can be understood as noise margin is different for each sampling plan and there exists an optimal sampling method. We tested several sampling methods that belong to symmetric group (translation, inversion, rotation symmetric), which are known to show better noise margin. The criteria to select the best sampling method were residual and linear term reproducibility which are significantly affected by raw data noise. The raw data variations include stage position errors and process induced alignment signal abnormality. We found among the candidates the optimal sampling method which leaves the least residual and shows as good repeatability as full chip measurement. Similar results could be obtained for overlay sampling method.

KEYWORDS: lithography, overlay, global alignment, sampling method

1. INTRODUCTION

There have been much efforts to improve the overlay control in the semiconductor industry. The lack of overlay control usually leads to the yield loss, where yield implies the percentage of electrically working chips, and the lowered device reliability. Thus the study to reduce the overlapping errors during exposure has attracted much attention among scanner providers and semiconductor manufacturers during past years. From the previous studies, which tried to reveal the major contributors to overlay control, global alignment noise (GAN) seems to be responsible for most overlay problem except for the case where the alignment marks are severely deformed due to fab. processes (e.g. CMP) and thus giving false offset. GAN is known to be generated by stage control inaccuracy, optical, electrical noise and process. Measurement position uncertainty due to GAN gives slightly incorrect overlay modeling results, which give rise to wafer-to-wafer overlay repeatability problem. These findings show that the overlay control would be improved by reducing GAN. Many improvements have been done with respect to stage control and alignment system. It is widely believed that the overlay control less than 30nm will be fulfilled within near future through the efforts to reduce GAN.

Considering the fact that most part of GAN is the problem of mechanical and electrical noises, which scanner providers have been trying to improve, but still remains in some degree, it seems to be true that there are no contributions from scanner users. However, recent studies[1] on sampling methods reveal that additional modification of overlay modeling can be found when different sampling methods applied. It is thus presumable that GAN behaves better when an optimized sampling method is applied. In this paper, we have argued that GAN margin can be improved by optimizing sampling methods. The better GAN margin implies that the wafer-to-wafer overlay modeling deviations is reduced in case of using the optimized sampling method. For this purpose, we first present that GAN can be varied according to process condition, machine ID, and mark design. Then it is shown that GAN affects the wafer-to-wafer overlay results. Finally the sampling method dependency of GAN and overlay and optimized sampling method are presented and discussed.

In *Metrology, Inspection, and Process Control for Microlithography XIV*, Neal T. Sullivan, Editor,
Proceedings of SPIE Vol. 3998 (2000) • 0277-786X/00/$15.00

2. BACKGROUND

Misalignment distributions on a wafer can be least-square-fitted with the linear overlay model. The main purpose of overlay modeling is to extract machine-correctable terms from the random distributions to minimize overlap errors between two layers. The rotation, skew, magnification and translation are well-known machine-correctable terms and, in case of SVGL Micrascan series, these can be given by the first-order coefficients of wafer cartesian coordinates in the following SVGL generic equations.

Interfield Equations (Grid / Wafer Coordinate System: [x,y]):

$$\begin{cases} \partial e_x = XTran + GXMag * x - GRot * y \\ \partial e_y = YTran + GYMag * y + (GRot + GSkew) * x \end{cases} \quad (1)$$

Intrafield Equations (Field Coordinate System: [x',y']):

$$\begin{cases} \partial e'_x = XRTran + (FXMag + FIMag) * x' - FRot * y' \\ \partial e'_y = YRTran + FIMag * y' + (FRot + FSkew) * x' \end{cases} \quad (2)$$

These coefficient calculations are performed for all every global alignment and overlay measurement. The global alignment reveals information regarding overlay errors to an exposure system before exposing the wafer. On the other hand, the overlay measurement reveals information regarding the exposed wafer overlay error to an operator. When the exposure system attempts to perform the global alignment, the correction terms from the overlay measurement are added to the global alignment result. In other words, the correction coefficients revealed by overlay measurements allow fine modifications of global alignment coefficients. Therefore, the coefficient error occurring at the global alignment or at the overlay measurement leads to undercorrected linear error. Even when all linear errors are corrected, high-order error (i.e., residual) remains. The residual, which is assumed to come from stage imperfections and process-generated non-uniformity with high spatial frequency, cannot be corrected systematically in the same manner as linear coefficients.

Considering eq. (1), one would expect that there need only six measurements to obtain six unknowns. However, there are more than six measurements in a wafer, which gives rise to the ambiguities of the coefficients. Furthermore, existing noisy measurements and residual, if both affect each other in the manner of amplifying total error, may cause unwanted fitting results. These errors can be reduced if one applies the sampling method that shows a large noise margin.

2.1 SAMPLING METHOD CHARACTERIZATION

2.1.1 NOISE ROBUSTNESS

The main causes of measurement position deviation (MPD) are known to be the limitation of stage positioning accuracy and the finite repeatability of alignment signal generation and detection. MPD generally results in the global alignment repeatability problem. One example showing how MPD affects the alignment result is shown in Fig. 1. The uncertainty of the aligned position or overlay measurement induces correction-term deviation. From the simple calculation, one can expect that the deviation will vary linearly according to the given dimensions, L and S. Figure 1 shows that calculated XMag and YMag deviation grows larger at small L and S. In other words, field-term repeatability is considered to depend on the arrangement of the measurement points. This suggests that for every real exposure system with finite MPD, the sampling method affects global alignment repeatability. Therefore one can expect that the optimized noise (i.e., MPD) robust sampling method, if found, will guarantee good overlay results.

2.1.2 RESIDUAL

When the displacements of every field in a wafer are measured and fitted by eq. (1), the resulting residual errors can be regarded as stage- or process-induced error. On the other hand, when a sparse sampling method is applied, it is probable that undercorrected correctable errors can be added to the residual, which results in a large residual error. Therefore, a sampling

mark usually reduces the mark signal strength or induces slight mark signal asymmetry. In some cases, one can find a noticeable difference of GAN level induced by lowered signal strength. Poly-silicon, one of the common materials in the fab., shows strong absorptance at alignment light wavelength band. With proper annealing treatment, however, the absorption is dramatically lowered due to increased degree of crystallinity of poly-silicon. To find out the GAN difference caused by the presence of annealing treatment, we prepared two poly-silicon deposited wafers, where one received annealing treatment and the other did not, and performed the repeated global alignments for each wafer (see Fig. 2-(a)). The higher GAN level for normal wafer, i.e. without annealing treatment, can be found.

3.1.2. ELECTRICAL AND MECHANICAL NOISE

Every exposure system is equipped with more than one mechanical stages. Despite lots of improvements made during past years, stage control accuracy remains in about 5 nm region. Thus the position information given by the alignment marks has some degree of uncertainties. In addition, there exists an extra uncertainty, which comes from alignment signal processing. We performed the experiment to view the noise level difference between two exposure systems. Under the condition of using a same product wafer and keeping an identical arrangement of measuring fields, clear GAN difference between two systems could be found (see Fig. 2-(b)).

3.1.3. *MARK DESIGN*

Most alignment systems prepare extra mark designs to deal with fatal processes. These optional mark designs maintain the fundamental dimension, a necessary condition for detection, but also have additional features. SVGL Micrascan series use both two-line and variable-box type features for global fine alignment (see Fig. 3-(a)). Although these two marks have noticeably different appearance, there is no problem to detect each mark. We prepared a wafer that has four different types

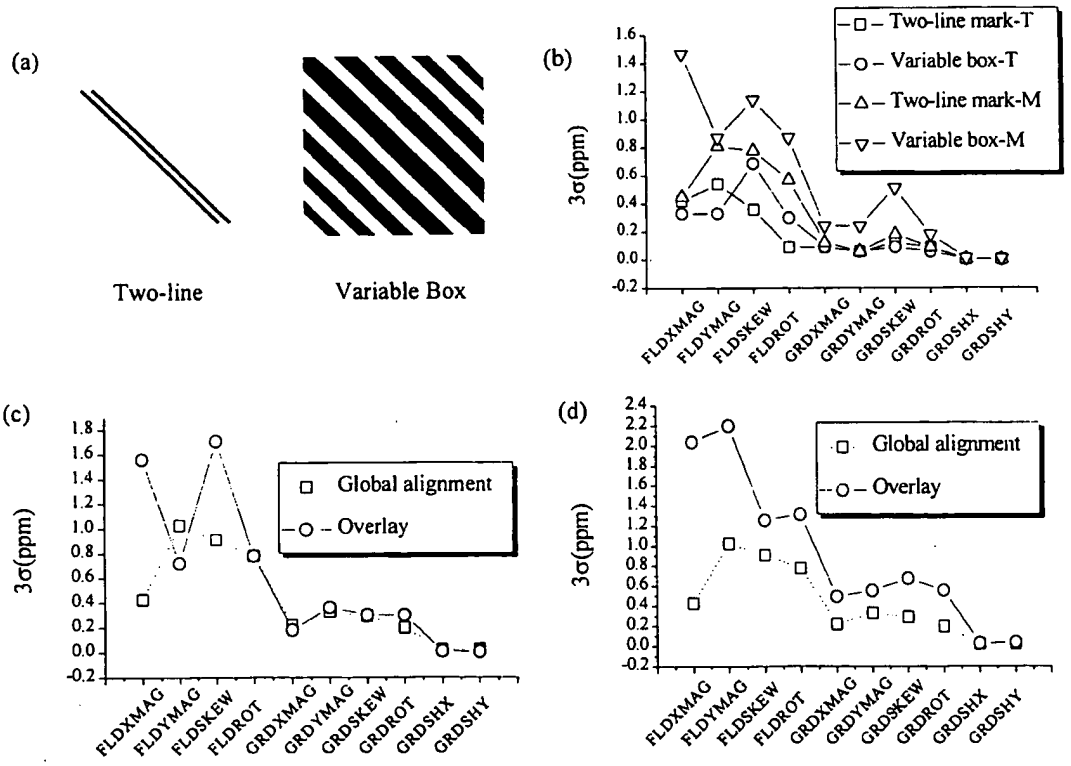


Figure 3. (a) Generic global alignment mark designs for SVGL Micrascan. (b) Global alignmen repeatability difference measured among different mark designs. (c) Comparison between globa alignment and wafer-to-wafer overlay deviations measured at front-end layer. (d) Back-end layer.

method that leaves the least residual might be the sampling method that we are seeking.

2.1.3 LINEAR COEFFICIENT

The difference in correctable terms measured between the reference plan, where every field in a wafer is measured, and the given sampling method, provides sampling method contributions to the total overlay errors. Therefore, when the given sampling method shows less difference than the others, one can consider this plan to be a good sampling method.

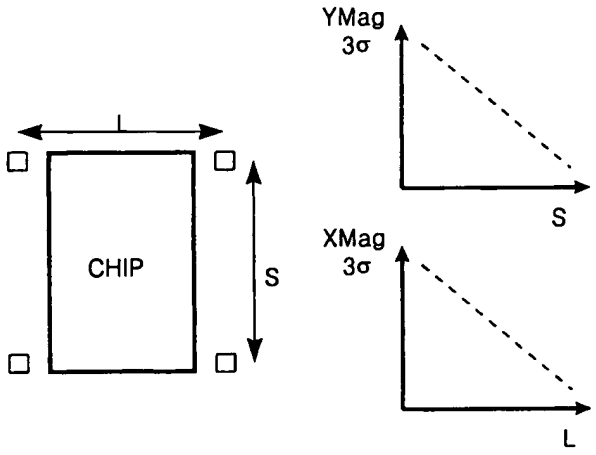


Figure 1. In-chip sampling position dependency of field correction terms.

3. RESULTS AND DISCUSSION

As discussed in the previous part, GAN is unavoidable in some degree. It is thus obvious that the repeated global alignment and overlay modeling for the same product wafer always shows a gaussian-like distribution centered at a non-zero model parameter with finite standard deviation. We will focus on the standard deviation, which can be regarded as GAN.

3.1 GLOBAL ALIGNMENT NOISE SOURCES

3.1.1. PROCESS

There are several processes that ought to be considered in the fab. The optically opaque film deposited over alignment

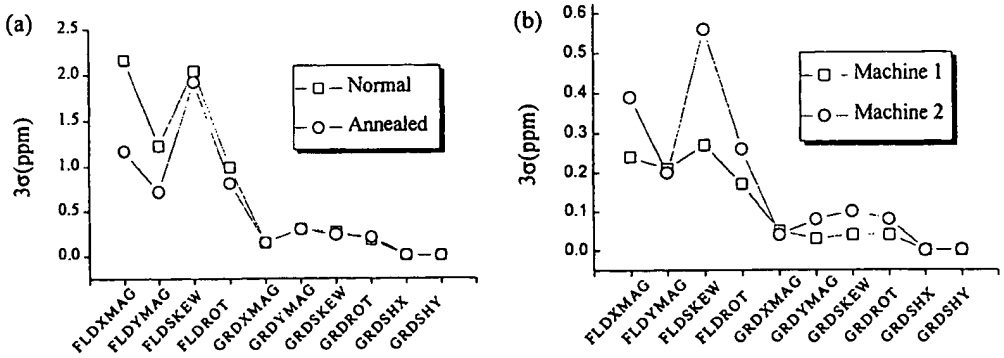


Figure 2. (a) Global alignment repeatability difference between a normal poly wafer and a annealed poly wafer. (b) GAN measured at different exposure tools.

of mark; a trench two-line, a mesa two-line, a trench variable box, and a mesa variable box. To obtain GAN level for each mark, 20-times repeated global alignments were performed respectively. In spite of keeping identical conditions except for different mark design applied, we were able to observe the different GAN level (see Fig. 3-(b)). Such differences can be understood as each mark has its own signal strength and process robustness. This observation leads to the well-known conclusion that GAN margin can be improved by optimizing mark type.

3.1.4. GLOBAL ALIGNMENT NOISE AND OVERLAY

Non-zero GAN implies that every wafer in a same lot experiences different overlay modeling respectively, therefore, the higher GAN leads to the worse wafer-to-wafer overlay. Figure 3-(c) and (d) show the comparison between the GAN of a wafer and wafer-to-wafer overlay. Y-axis represents the three sigma of repeated global alignment and wafer-to-wafer overlay respectively. The plots shown in figure 3-(c) and (d) correspond to a wafer selected from gate stack process and a wafer from storage node process, respectively. From these plots, one can observe that GAN levels of each wafer are similar but wafer-to-wafer overlay of each wafer exhibit different results. The wafer from gate stack process shows GAN limited overlay behavior, i.e., wafer-to-wafer overlay deviation is less than GAN level. On the other hand, the wafer from storage node process shows that wafer-to-wafer overlay deviation seems not to be limited by GAN but by some other noises, such as, the errors from mark deformation.

3.2 OVERLAY SAMPLING METHOD

Measuring the overlapping error between two layers requires several box-in-box marks to identify the error. For most cases, these features are located at four symmetrical corners of the field. Overlapping errors are measured from each marks and an error vector representing that field can be obtained by sum of four error vectors. From overlay model fit of error vectors, one can calculate interfield corrections corresponding to a given sample plan. These interfield corrections are considered to contain the slight error due to the sampling. In order to minimize the error due to the overlay sample plan, we proposed six sample plans which maintain the rotation symmetric arrangement, which is based on the assumption[2] that the symmetric arrangement provides better results than the asymmetric one. Each plan is designed in order to clearly identify

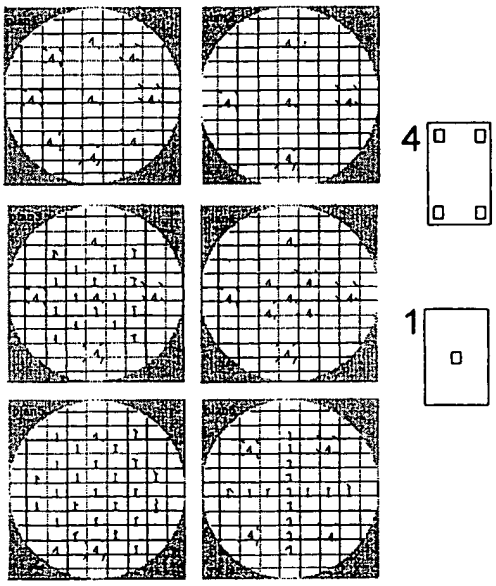


Figure 4. Overlay sampling methods proposed.

the arrangement characteristics and simultaneously require the same time to complete measurement. In Fig. 4, which shows the six candidates with its own typical field arrangement, one can observe that plan 1 and 2 have weak correlation among measuring fields maintaining rotation symmetry while plan 4 shows strong correlation among central fields, and that the other plans have strong correlation and good coverage. To allocate more grid sample points over a wafer and simultaneously to satisfy the requirement that they must take the same length of time to measure, there are some fields where only one measurement occurs per field. One can find this field in plans 3 to 6. We indicate this field by "1", while the normal field by "4".

We performed linear-term repeatability tests for these six sample plans. The wafers prepared for these experiments were carefully selected from real product wafers. The exposure tool for patterning wafers was the KrF Scanner, SVGL Micrascan III. To exclude the stack layer contribution from overlay results, we selected wafers from different layer stack conditions, which include gate stack, direct contact stack, and storage node stack substructures. For each wafer, we measured overlay distributions for six different sampling plans and calculated the residual and linear terms. Figure 5 shows the residual and

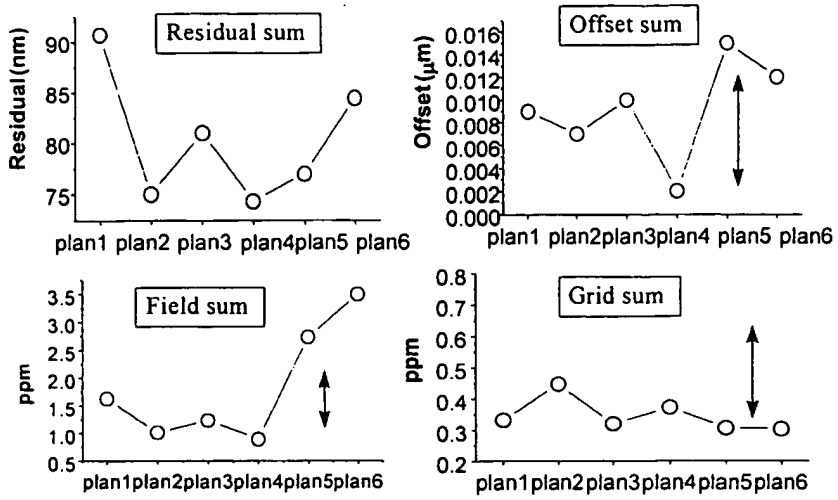


Figure 5. Residual, offset, field term and grid term difference between each pla and reference plan (every chip in a wafer are measured).

linear term deviations for each sample plan. These plots can be regarded as sufficiently reliable to provide information on the optimum sample plan because each data point is summed over three different wafers where both overlay distributions and noise levels differ from each other. Residual sums, offset sums, and field sums reveal that plan 4 is more noise robust and leaves less residual than the other plans. Grid sums seem to yield ambiguous result. However, if one considers the significant scale of grid terms, it is understandable that plan 4 works as well as the other plans. These results can be understood in the sense that the plan of higher coverage over the wafer map yields more correct grid terms and that the plan, in which measuring fields are closely placed, exhibits good performance in terms of giving correct field terms, offsets, and residual.

3.3 GLOBAL ALIGNMENT SAMPLING METHOD

We performed an experiment that reveals the relation between global alignment repeatability and the wafer-to-wafer overlay deviation. For this evaluation, the overlay of 25 wafers and the 20-times-repeated global alignment of a wafer are

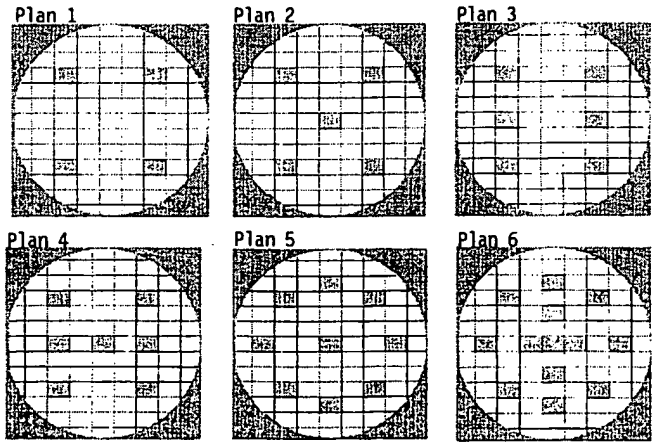


Figure 6. Global alignment sampling methods proposed.

measured. These wafers are photoresist-coated product wafers and global alignment is performed on the SVGL standard inclined two-line target. Then, we calculated the standard deviations of the linear terms. It should be noted that the different statistics were applied to the wafer-to-wafer overlay and the single wafer global alignment. In order to optimize the global alignment sample plan, we considered the linear-term repeatability and coordinate matching between the global alignment

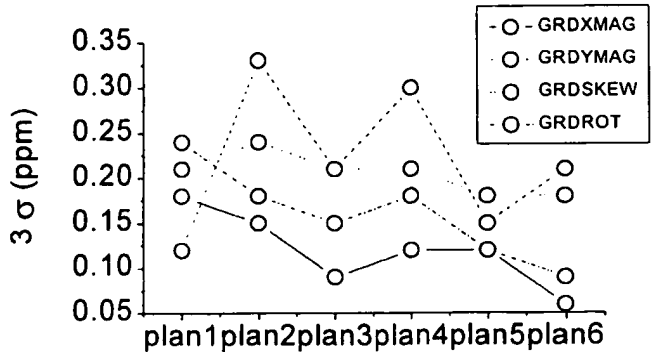


Figure 7. Global alignment repeatability difference among sampling methods.

and overlay measurement. If the selected fields for global alignment use do not coincide with the locations of the fields for the overlay measurement, the raw displacement information for the calculation differs, and as a result, overlay error may occur due to different calculations. In addition to this concern, from eq. (1), it is obvious from mathematical point of view that the coordinates of global alignment and overlay measurement had better coincide with. With the restriction of the measuring field arrangement, we performed global alignment repeatability measurements for six sample plans (see Fig. 6). Each plan is mainly distinguished by the number of global alignment chips, ranging from 4 to 13, while all plans follow the reference arrangement based on the overlay sample plan. We performed global alignment 20 times with respect to each sample plan. Figure 7 shows the three σ's of the alignment statistics over 20 global alignments, in which grid rotation, grid mag, and grid skew are plotted. Interestingly, although we maintained the similar arrangements and therefore expected that the three σ's would decrease linearly with increasing shot number, however, we found that the repeatability behaves sensitively to a slight change in the shot arrangement. On the other hand, long-range behavior follows our expectation that the three σ's would decrease as the number of shots increases. These observations can be understood as noise robustness can be improved not only by the number of shots for global alignment use, but by the optimized arrangement of the shots.

As a concluding remark, the clear repeatability difference among the sample plans is evident and plan 5 seems to exhibit better noise robustness than plan 6.

4. CONCLUSION

We investigated the possible causes that yield global alignment noise and the sample plan dependency of overlay and alignment control. In order to optimize the sample plan, we first proposed several candidates based on symmetry considerations and coordinate matching between global alignment and overlay measurement, and then performed systematic tests to determine the difference between the plans. The residual, correctable-term accuracy, and noise robustness were found to function well as criteria. The new approach to the optimization of the sample plans revealed that the specific candidate performs better than the others. We hope that this result emphasizes the importance of sample plans and that our novel approach could aid.

REFERENCES

1. J.S. Hong, J.H. Lee, J.S. Park, H.K. Cho and J.T. Moon, "Optimization of sample plan for overlay and alignment accuracy improvement", Jpn. J. Appl. Phys. **38**, pp. 7164-7167 (1999).
2. I.D. Fink, N.T.Sullivan and J.S.Lekas, "Overlay sample plan optimization for the detection of higher order contributions to misalignment", Proc. SPIE **2196**, pp. 389-399 (1994).

Overlay Simulator for Wafer Steppers

William H. Arnold

Advanced Micro Devices, Inc.
901 Thompson Place, MS 79
Sunnyvale, CA 94088

Abstract

The impact of wafer stepper overlay errors on device yields and design rules are studied. First, the classical Lynch model for normally distributed sizing and overlay errors is reformulated for orthogonal geometries. Then the distribution of overlay errors in the linear Perloff model describing global alignment is derived. Finally, a Monte Carlo program, OVS, for simulating stepper overlay errors is introduced. OVS is used to determine the impact of individual component errors, such as those due to lens distortion or to mask making, on the overall distribution of errors.

Introduction

Pattern overlay tolerance is one of the key design rules for integrated circuit manufacturing. The current method of specifying overlay tolerances is to state, on the basis of simple error budgets, the value of the overlay error at a given high percentage of the total measured data. There are at least three major problems with this[1]. First, most error budgets treat all errors as random when in fact some are systematic. Second, these error budgets rarely take field size or wafer size into account. Finally, rarely are enough data points taken to ensure what is really important: that no point within the imaged array of chips be misregistered by more than the design rule.

An overlay simulation computer program, OVS, has been written to predict the total distribution of errors given as input the characteristic component error distributions in interfield errors (translation, wafer rotation and expansion, orthogonality) and intrafield errors (die rotation, magnification, trapezoid, and distortion). Reticle to reticle stacking errors are taken into account. These are used in a Monte Carlo simulation to generate the expected total error distribution over thousands of individual fields and wafers. The program can be used to predict the performance of a single stepper or of a large group of randomly mixed steppers.

The underlying model for the simulation is the linear interfield model introduced by Perloff[2] combined with the polynomial intrafield model introduced by MacMillen and Ryden[3]. Each error is given a systematic offset and a random component characterized by a Gaussian distribution of a given sigma. The simulation computes the errors found at F points per field, and W points per wafer, where the number of points and their locations are set by the user. If any point within the field exceeds the overlay tolerance it is noted by the program and the field is termed a "bad" field.

The fit of simulation data to experimental data is found to be quite good. One interesting result of the simulations is that many distributions are platykurtic, i.e., have less data in the tails of the distribution than the best fit Gaussian computed from the same data, a tendency noticed by many workers from experimental data. It is also found that as the size of systematic errors grow with respect to random errors, the more the distributions tend to be platykurtic.

Emphasis is placed in this report on determining the overall distribution of overlay errors given the distribution of component errors. First considered is a simple overlay model proposed by Lynch[4] which will allow development of the ideas advanced here against a classical background. Then consequences of the linear model[2] will be explored in terms of its predictions for global overlay error distributions. It is found, for example, that the linear model predicts a semicircular histogram of overlay errors on a given wafer aligned at two points. In the last section the overlay simulator model and details of the computer program are discussed. The effects of individual component errors on the entire error distribution are examined, assuming that the distributions of component errors can be measured and specified[5].

As a result of this work, a new proposal is made for the specification of overlay errors which corresponds more closely to the desires of chip designers. The proposal is called the "good fields rule", in which stepper overlay is quantified in terms of the percentage of individual fields which contain no error greater than the specification.

The good fields rule for total overlay satsfies many intuitions about how overlay affects chip yields and is more stringent than any presently offered by stepper manufacturers. While a large number of points within the usable image field may be properly registered, it only takes one bad point to cause circuit failure, resulting in complete yield loss on single die reticles and fractional yield on multidie reticles.

Lynch Model

Lynch[4] studied the problem of contact window limited yields for LSI devices. In his analysis, the process yield for defining and aligning a circular contact window over a larger circular pad of polysilicon is derived. Yield is defined as the contact window falling completely on the poly pad. The minimum contact window size is set by the aligner's resolution capability. The required size of the poly pad to achieve a given yield is then calculated from contact size, the process standard deviations for edge control for both the contact and the poly pad, and the contact to poly alignment standard deviation. Lynch's model assumes explicitly that the distributions of edge controls for contact and poly as well as the alignment errors are Gaussian.

Device designers typically do not set radial design rules but instead set rules along orthogonal directions. Thus a more germane problem is that of a square contact window falling on a larger square polysilicon feature. If we assume in this case that there is no asymmetry between edge placement or alignment precision in the X and Y directions then Lynch's expression can be recalculated. The geometry of the problem is shown in Figures 1a and 1b. The contact window is free to expand or contract with mean size C and standard deviation s_c. The poly pad likewise has mean size L and standard deviation s_1. The overlay error is assumed to have zero mean error and a standard deviation along the X or Y directions of s_0. Normal distributions are assumed for each.

The problem of the lithographer is to specify the smallest tolerance, $T = 1/2(L - C)$, which will allow the contact to fall completely on the poly pad at a given high percentage level. The tolerance is the nominal distance between the edge of the contact to the nearest poly edge. In a given case, this distance will change to $D = 1/2(l - c)$ where l is the true size of the pad and c the true size of the window. The distribution of D is also Gaussian[6] with mean T and standard deviation $(s_c^2 + s_1^2)^{1/2}$

$$f(D) = (2\pi(s_c^2+s_1^2))^{-1/2} \quad \exp(-(T - D)^2/ 2(s_c^2+ s_1^2)) \qquad (1)$$

The probability that the contact falls completely on poly is then the probability that $D - |dx|$ and $D - |dy| > 0$. The probability that $|dx|$ (or $|dy|$) $< D$ for a given D is given by

$$P(|dx|, |dy| < D) = (1 - 2\bar{Q}(D / s_0))^2 \quad \text{where } \bar{Q}(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} \exp(-t^2/2)dt \qquad (2)$$

The probability then that the contact falls on the poly, that it yields at a given tolerance T, is then the convolution of eqn. 1 and eqn. 2:

$$Y(T) = (2\pi(s_c^2+s_1^2))^{-1/2} \int_{0}^{\infty} (1 - 2\bar{Q}(D/s_0))^2 \exp(-(T-D)^2 / 2(s_c^2+s_1^2)) \, dD \qquad (3)$$

This expression allows one to find how big to make the tolerance given normally distributed process variations in poly and contact sizing, and in overlay. As might be expected, it satisfies the root sum square error budget calculation typically used[7] to calculate the tolerance. This can be seen in Figure 2 in which Y(T) is plotted as a function of T for three different values of overlay precision. The contact sizing precision is 0.05 μm, poly precision is 0.03 μm, and the three overlay precisions are 0.05, 0.1, and 0.15 μm, all numbers at one standard deviation. For the 0.15 μm, one sigma case the root sum square of the sizing and overlay precisions is 0.22 μm. Y(T) = 0.68 at 0.22 μm, which is the percentage included within one standard deviation in a Gaussian or normal distribution, demonstrating the underlying consistency of the Lynch model.

From these simple considerations it can be seen clearly why overlay is such a key parameter for lithography. For example, assume that the smallest contact size which can be reliably defined with 0.05 μm, 1 sigma control is 1.0 μm. The poly width $L = C + 2T$ for this case is 1.48 μm at the 99.7% yield level, and 2.00 μm for the case where overlay control is 0.15 μm, one sigma. The area of the poly pad can be cut by 45% by increasing overlay precision from 0.45 μm, 3 sigma, to 0.15 μm, 3 sigma. Since a CMOS transistor cell size is proportional to this area, it can be seen how sensitive device size is to overlay precision.

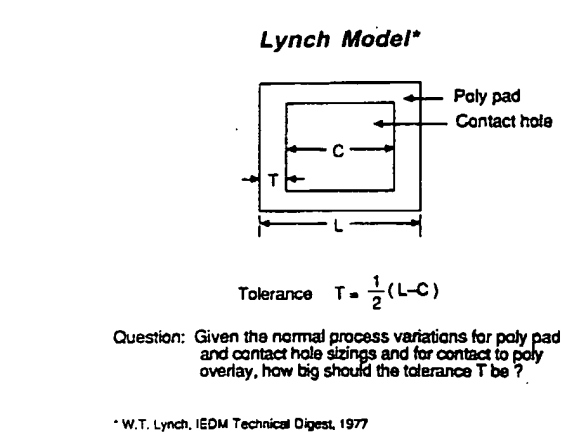


Figure 1a. Definition of the tolerance T in the Lynch problem.

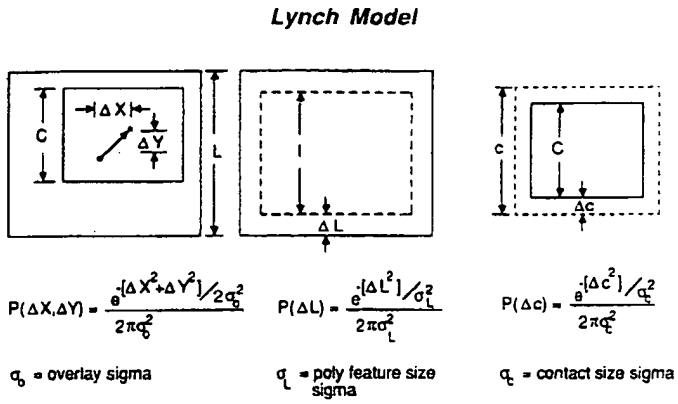


Figure 1b. Probability distributions for contact window and poly pad sizings, and for contact to poly alignment.

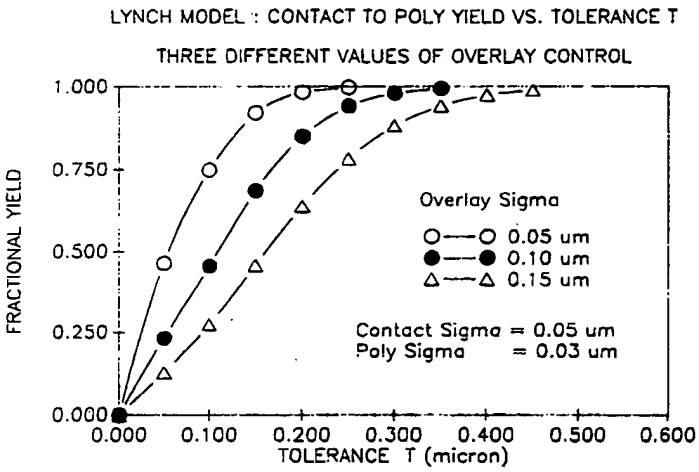


Figure 2. Fra ctional contact to poly yield versus the tolerance T for three different values of overlay precision.

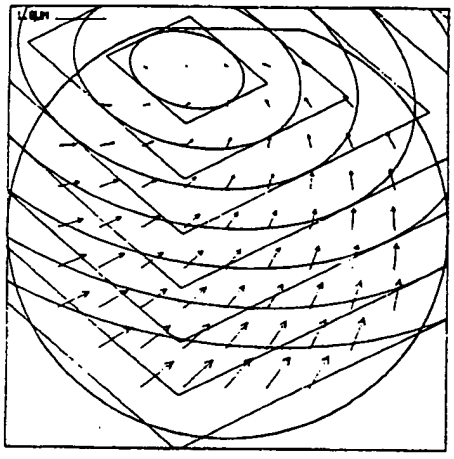


Figure 4. Example of the contour representation of linear overlay errors. T_x = .19 μm, T_y = .30 μm, θ_x = 7.3 ppm, θ_y = 3.9 ppm, E_x = 6.5 ppm, E_y = - 7.5 ppm.

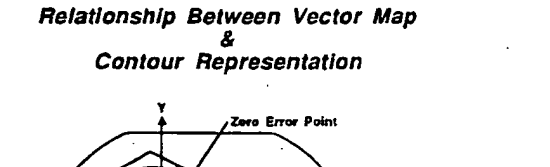


Figure 3. Relationship between the vector and contour representations of linear overlay errors.

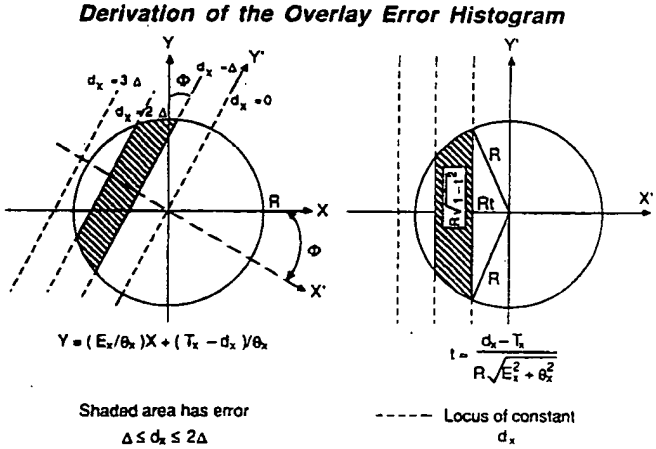


Figure 5. Derivation of the linear overlay error histogram by sectioning the circle and determining the relative area between successive equal error contrours.

Overlay Error Distribution for Global Alignment

Interfield registration errors made by reduction steppers using two-point global alignment can be modelled successfully by linear regression analysis as shown by Perloff and co-workers[2,8]. Kim and Ham[9] showed that assumption of the linear model of registration errors leads to the result that the loci of equal-valued errors in the Cartesian coordinates x and y on the wafer are straight lines which intersect at oblique angles. There is a single point in the xy (wafer) plane at which the registration error is zero, which can fall either on the wafer or outside it. Errors less than a constant absolute value fall within a parallelogram centered on the zero error point. Equal valued vector displacements have elliptical contours.

Registration errors on a globally aligned wafer are estimated independently to first order in the Cartesian coordinates x and y as

$$dx = T_x - \theta_x y + E_x x + s_x \tag{4}$$

$$dy = T_y + \theta_y x + E_y y + s_y \tag{5}$$

where T_x, T_y, E_x, E_y, θ_x, and θ_y are parameters formed by linear regression of a data set consisting of N sets of errors (dx, dy) at wafer locations (x, y). The s_x and s_y are residual errors not fitted and are usually associated with random stage stepping errors. The first order error parameters can be grouped into three simple geometrical classes : translation (T_x, T_y), rotation and orthogonality (θ_x, θ_y), and expansion (E_x, E_y).

The distribution of errors across a wafer predicted by the linear model can be determined by exploring the behavior of eqn (4) and (5) at different wafer locations. One interesting result shown by Kim and Ham is that the loci of equal-valued errors in x and y are straight lines. This can be seen by solving eqn (4) for y (assume $s_x = s_y = 0$) :

$$y = (E_x/\theta_x)x + (T_x - dx)/\theta_x \tag{6}$$

which is of the linear form $y = ax + b$, where the slope $a = (E_x/\theta_x)$ and the y intercept $b = (T_x - dx)/\theta_x$. Similarly, solution of eqn (5) yields $y = -(\theta_y/E_y)\,x + (T_y + dy)/E_y$. Simultaneous solution of eqns. (4) and (5) when $dx = dy = 0$ yields the position of the zero error point:

$$x_o = -\frac{(T_x E_y + \theta_x T_y)}{\theta_x\theta_y + E_x E_y} \; ; \quad y_o = -\frac{(E_x T_y - \theta_y T_x)}{\theta_x\theta_y + E_x E_y} \tag{7}$$

Depending on the wafer radius, the origin of coordinates and the magnitude of the errors in T, θ, and E, the zero error point can either fall on the wafer or outside it (i.e., there is no point on the wafer with perfect overlay). A simple interpretation is that there is a single point in the xy plane at which there is zero overlay error and that error in x and y less than $dx = \pm c$ and $dy = \pm d$ respectively are bounded by a parallelogram with sides described by

$$y = (E_x/\theta_x)x + (T_x \mp c)/\theta_x \tag{8}$$

$$y = -(\theta_y/E_y)x + (T_y \pm d)/E_y \tag{9}$$

The parallelogram is centered on the zero error point (x_o, y_o). This is illustrated in Figure 3. Lines of constant dx are inclined at an angle Φ with respect to the x axis, while lines of constant dy are inclined at an angle Ψ where $\tan\Phi = E_x/\theta_x$ and $\tan\Psi = -E_y/\theta_y$.

If one instead considers the locus of equal valued vector errors ($d_r^2 = (d_x^2 + d_y^2)^{1/2}$), then the contours are ellipses, also centered on the zero error point, with the semi-major axis inclined at an angle Φ_r with respect to the x axis:

$$\tan 2\Phi_r = \frac{2(\theta_y E_y - \theta_x E_x)}{(E_x^2+\theta_x^2)-(E_y^2+\theta_y^2)} \tag{10}$$

A simple illustration of these results is the special case of isotropic expansion error without translation or rotation. Traditional vector maps represent this type of error with radially directed arrows whose lengths grow as the distance from the center increases. This method suggests that the same error can be represented with circular contours, where the circles are centered at the origin and whose radii grow linearly with increasing distance from the origin. Contours of constant and equal dx and dy are squares. The case of pure rotation error without translation or expansion also yields

circular contours centered at the origin. The general case, where translation, rotation, and expansion errors are present, yields ellipses and parallelograms centered on the zero error point which is not the origin of coordinates. Figure 4 shows an example wafer for which both the vector map and contour representations are given. The error parameters are listed in the figure.

What relationship exists between the model parameters and the distribution of overlay errors on a single wafer? Since one can compute the straight line contours of constant dx or dy increments, imagine sectioning the wafer into areas between successive dx or dy increments (see Figure 5). Once this is done the relative size of each area can be calculated using standard relations for the sector of a circle which can then be used to construct the histogram of errors on the wafer. The mathematical details of the derivation are given in reference 12.

The histogram of x overlay errors, H(dx), on a wafer of radius R and with error parameters T_x, θ_x, and E_x is given by

$$H(dx) = \frac{2}{\pi} \frac{d}{R(\theta_x^2+E_x^2)^{1/2}} \left[1 - \frac{(dx - T_x)^2}{R^2(\theta_x^2+E_x^2)}\right]^{1/2} \qquad (11)$$

where H(dx) represents the percentage of wafer area within $\pm$ d/2 of dx and dx ranges between $T_x - R(\theta_x^2+E_x^2)^{1/2}$ and $T_x + R(\theta_x^2+E_x^2)^{1/2}$. The expression for the y histogram, H(dy), is identical to this with all subscripts x replaced by y.

This result explains why overlay errors across a single wafer which is aligned at two points are in general not normally distributed. The form of the histogram is semicircular! It is straightforward to calculate that the standard deviation of this distribution is $s = (1/2)R(\theta_x^2+E_x^2)^{1/2}$. Thus the entire distribution is contained within $\pm$ 2s of the mean T_x.

Figures 6a and 6b show the measured and modeled X axis overlay errors for the example wafer illustrated in Figure 4. The modeled histogram was calculated using equation 11. Random stage errors account for most of the difference between the modeled and measured data.

OVS : Overlay Simulator for Wafer Steppers

Anyone who has studied wafer stepper overlay specifications knows that it is a difficult job to translate the stepper vendor's specifications into numbers which device designers can use. The reasons are that the vendor specification usually refers to an ideal test case, in which wafers with nearly perfect alignment targets are used. Field size is usually restricted to values less than the maximum field. Only a limited number of points across the image field and across the wafer are actually measured. The vendor then guarantees that a given percentage of the total number of measurements will fall below the maximum specified overlay error. For example, one major stepper vendor specifies that overlay be measured at 17 points per image field and at 17 separate fields on each of three wafers. This gives a total of 17 X 17 X 3 = 867 data points per axis. Only 9 points per axis are allowed to exceed the overlay specification.

This type of sampling plan, however, does not guarantee that chips designed with an overlay tolerance equal to the vendor specification will yield at the same high percentage. The major reason is that it only takes one bad point, i.e., one location where the overlay rule is violated, to cause circuit failure and zero yield for that chip. It doesn't matter that 99% or more of the rest of the chip's area is overlaid within specification, the chip still doesn't yield. In the example sampling plan described, assume that matched steppers are used to print only one die per field. In the worst case, one might have a bad corner of the field due to lens to lens distortion differences in which the x overlay error is very close to the spec limit without considering alignment errors. The sampling plan would allow this point to exceed specification in 9 fields out of 17 X 3 = 51 total. A similar situation in the y direction added to this could lead to 18 fields out of 51 containing a point which violated the design rule. The vendor's specification that 99% of the errors were less than the given tolerance is met, yet only about 65% of the chips printed would yield.

One response to this reasoning is that such a situation is highly unlikely. Unfortunately, it is nearly as unlikely that only one field would contain all the overlay errors, so that the yield of image fields without bad points would in general be less than 98% (50 of 51). It must be remembered that modern devices have 15 or more mask layers, 5 or so of which require the tightest overlay specification, so that total overlay yield is roughly the yield of one alignment raised to the power 5 (or greater). Even at 98% single layer yield there is only 90% yield after five critical alignments and

82% after 10.

Given these considerations a new proposal is made for the specification of overlay errors which corresponds more closely to the desires of chip designers. The proposal is called the "good fields rule":

X% of all fields overlaid to a previous field will contain no bad points, i.e., errors greater than the specified overlay error.

X% can be a given high percentage, e.g., 99% or 99.7%, depending on device yield considerations. The good fields rule for total overlay satsfies many intuitions about how overlay affects chip yields and is more stringent than any presently offered by stepper manufacturers.

One serious logistics problem with the good fields rule is the relatively large number of measurements that need to be taken to ensure the specification directly. It is however possible to indirectly estimate the error distribution through computer simulation of the overlay process using as input the characteristic distributions of subcomponent errors such as reduction, die rotation, trapezoid, distortion, and so forth. To meet this need, an overlay simulation program, OVS, has been written. The program incorporates a Monte Carlo routine which simulates the alignment of many reticles and wafers and reports the statistics on all the errors found. The program can simulate the distribution of errors expected from a single stepper and from mixed steppers. Errors taken into account, assuming global alignment, are listed in Table 1.

Table 1 - OVS Input Error Parameters

Intrafield Errors - Single Stepper	
1) random magnification error*	M
2) random reticle rotation error*	R
3) random reticle stacking error*	s_1
Mixed Steppers	
4) relative distortion between two lenses i and j@	$D_1(i)-D_1(j)$
Interfield Errors	
1) random translation offsets in x and y*	T_x, T_y
2) random rotation offset*	θ
3) random orthogonality offset*	$d\theta$
4) random symmetrical expansion (or scaling) offset*	E
5) random assymetrical expansion offset*	dE
6) random variations around offsets for:! translation, rotation, and expansion	
7) random stage errors&	ss_k

Field locations l = 1 to f; wafer locations k = 1 to w

* - constant for one reticle change
! - chosen for each wafer alignment
@ - constant for all reticle changes
& - chosen at the center of each field k

The program starts by reading intrafield distortion errors for lenses i and j from an input data file. The relative intrafield distortion at 9 field locations is then calculated (in principle, any number of field points can be used). This data remains constant throughout the rest of the simulation. A single stepper is simulated by setting all the intrafield distortion errors in the data files to zero.

The program then starts a series of loops to simulate the errors incurred by the change of the reticle and random pressure and temperature variations between reticle changes. For each reticle change, random die rotation, magnification, and reticle stacking errors are chosen by the Monte Carlo routine. All component errors in this model except relative distortion between lenses are assumed to be Gaussian distributed[5]. A random number between 1 and 1663 is chosen by a random number generator. The computer then searches a look-up table which associates the random number with a t-value such that $-3 < t < 3$. The random magnification error is then calcualted as $M = ts_M$, where s_M is the input standard deviation for magnification control. In order for the simulation to be as accurate as possible it is necessary to have good information on the distributions of the component errors. Likewise, random values are chosen for reticle rotation. Reticle stacking errors

are generated at each of the 9 field locations.

Once the magnification, die rotation, and stacking errors are generated, the intrafield errors for the lot of wafers are calculated from (l = 1 to 9)

$$DX_l(i,j) = - R\, y_l + M\, x_l + (D_{lx}(i) - D_{lx}(j)) + s_{lx} \quad (12)$$

$$DY_l(i,j) = R\, x_l + M\, y_l + (D_{ly}(i) - D_{ly}(j)) + s_{ly} \quad (13)$$

The program then enters the wafer alignment loop. It first generates random offsets for x and y translation, wafer rotation, orthogonality ($d\theta = \theta_x - \theta_y$), scaling, and differential scaling ($dE = E_x - E_y$) for the wafer lot. For each wafer the program calculates a translation error in x by generating a random variation around the offset and adds that to the offset determined at the reticle change. The same is done for y translation, x and y rotation, and x and y scaling. The errors at the centers of the fields at all wafer locations k for that wafer are then calculated from

$$WX_k = T_x - \theta_x\, y_k + E_x\, x_k + ss_x \quad (14)$$

$$WY_k = T_y + \theta_y\, x_k + E_y\, y_k + ss_y \quad (15)$$

where ss_x and ss_y are random stage errors, generated at each field separately.

The total overlay error at each field location l and wafer location k is then calculated as the sum

$$VX_{kl} = DX_l + WX_k \quad (16)$$

$$VY_{kl} = DY_l + WY_k \quad (17)$$

For each wafer the VX_{kl} and VY_{kl} form two matrices. In order for a field to be considered good, the corresponding rows in VX_{kl} and VY_{kl} must contain no errors greater than the specification. The computer sorts through the matrices to find bad fields. It also calculates the mean x and y errors and x and y standard deviations for all matrices. Alignment is repeated as often as chosen. In many of the computer runs considered here, 10 wafers are aligned before the next reticle change.

Once the lot of wafers is finished, the computer changes the reticle again and repeats the procedure described above. After the computer runs through all the reticle changes requested, it then calculates the overall statistics for the run, including the mean and standard deviations in x and y for the entire distribution and the percentage good fields.

It is an easy modification to the program to simulate the behavior of steppers which employ field by field alignment rather than global alignment. In this case scaling, orthogonality, and random stage errors are set to zero. Translational and, if appropriate, die rotational alignment errors are generated at each wafer location k.

The program described here does not calculate nonlinear errors due to trapezoid, or due to third or fifth order distortion. Nonlinear intrafield errors are lumped together as unchanging signatures of lenses, and are characterized through the DX and DY input files. This is not an altogether justifiable assumption since trapezoid errors can change and be adjusted. It has also been shown that illuminator defocus can lead to third order intrafield errors which mimic true lens distortion[10]. However, these errors are assumed to be quite small in comparison with the others considered here. It is certainly possible to continue the analysis to cover the case of time varying nonlinear intrafield errors.

As an example of the use of OVS, two different types of 5X reduction wafer stepper were simulated. The first type uses off-axis global alignment and is uncompensated for magnification changes as a function of barometric pressure. The second type represents a stepper with through the lens alignment, barometric compensation, and the ability to adjust magnification to match wafer expansion due to processing. Global alignment is also used on this type stepper. In this example, single stepper performance is simulated by assuming that all intrafield errors except reticle rotation, magnification, and reticle stacking errors are zero.

Table 2 lists the input parameters for the two simulations. Table 3 lists the percentage good fields found at various overlay tolerances and the equivalent $\bar{X}$ + 3 sigma values. Figure 7 is a plot of good fields percentage versus the overlay design rule.

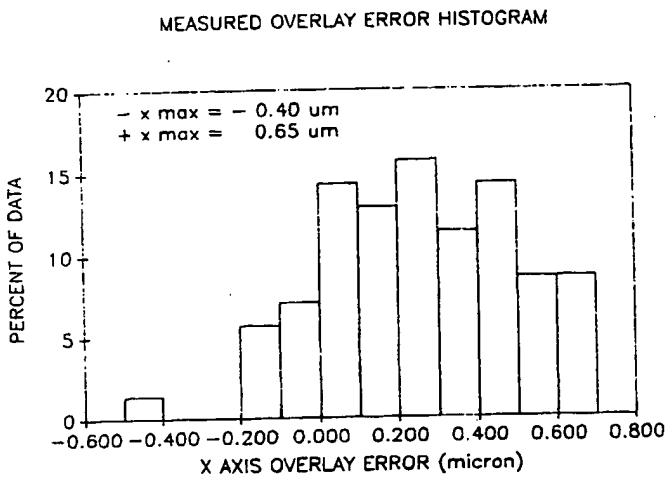


Figure 6a. Measured X overlay error histogram for the example wafer pictured in Figure 4.

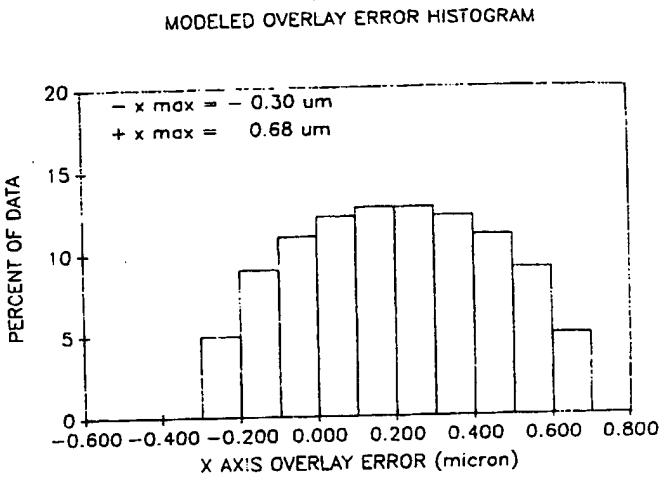


Figure 6b. Modeled X overlay error histogram for example wafer in Figure 4. Histogram calculated using equation (11).

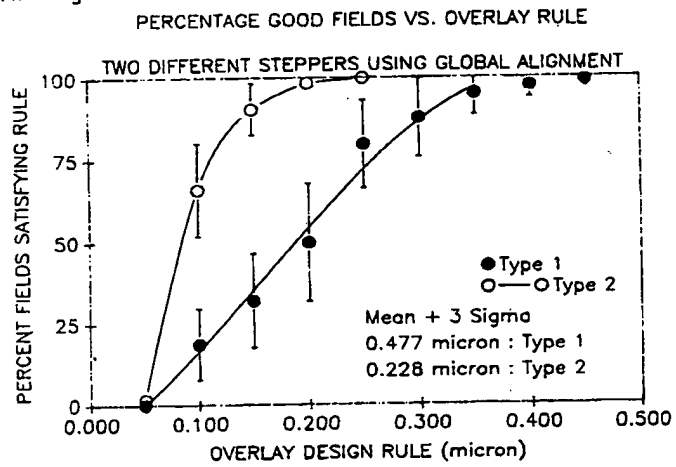


Figure 7. Percentage good fields versus the overlay design rule for two different types of stepper (see Tables 2 and 3).

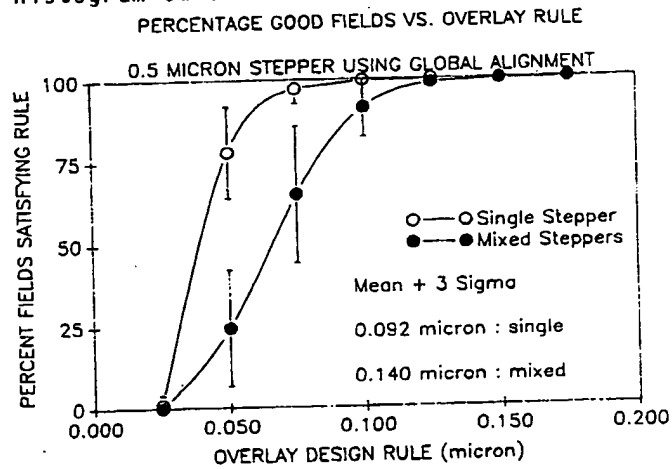


Figure 8. Percentage good fields versus the overlay design rule for a stepper using global alignment to achieve 0.1 μm overlay, single machine, 0.15 μm mixed.

PERCENTAGE GOOD FIELDS VS. OVERLAY RULE

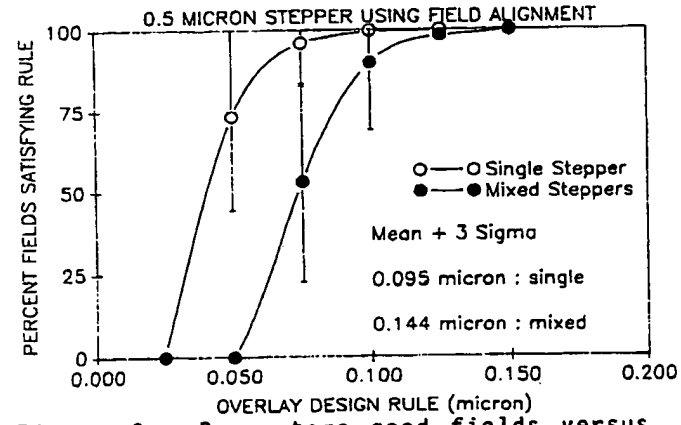
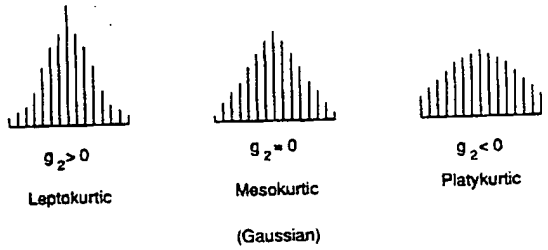


Figure 9. Percentage good fields versus the overlay design rule for a stepper using field alignment to achieve 0.1 μm overlay for single machines, 0.15 μm mixed steppers.

Coefficient of Kurtosis: g_2



Reference: ASTM STP 15D, Manual on the Presentation of Data and Control Chart Analysis

Table 2 - Characteristic Component Overlay Errors

Interfield Errors	Units	Stepper 1	Stepper 2
Translation offset sigma	µm	.07	.03
Translation sigma around offset	µm	.07	.03
Symmetrical expansion offset sigma	ppm	1.0	0.5
Asymmetrical expansion offset sigma	ppm	0.5	0.5
Expansion sigma around offset	ppm	1.0	0.5
Wafer rotation offset sigma	ppm	1.0	0.5
Orthogonality offset sigma	ppm	0.5	0.5
Rotation sigma around offset	ppm	1.5	0.5
Stage precision sigma	µm	.04	.03
Intrafield Errors			
Magnification sigma around offset	ppm	8.0	3.0
Reticle rotation sigma around offset	ppm	4.0	3.0
Reticle stacking error sigma (@ 1/5X)	µm	.02	.02

Offset sigma refers to run to run variation of the mean. Sigma around the offset refers to variation within the run. In the simulation a run is defined as the number of wafers aligned before the next reticle change.

Table 3 - Overlay Simulation Results

Design Rule (micron)	% Good Fields:Type 1 Mean	% Good Fields:Type 1 Std. Dev.	% Good Fields:Type 2 Mean	% Good Fields:Type 2 Std. Dev.
.05	0.1	0.1	1.7	2.5
.10	18.1	11.1	65.9	14.2
.15	21.7	14.5	90.5	7.9
.20	49.7	18.1	98.4	2.6
.25	79.7	13.2	99.8	0.6
.30	87.7	11.9	99.9	0.1
.35	95.1	6.7	100.0	0.0
.40	97.5	3.2		
.45	98.9	2.1		
.50	99.7	0.4		
.55	99.9	0.1		
.60	100.0	0.0		
$\bar{X}$ + 3 sigma calculated for all data (micron)	0.477		0.228	

Assumes 9 points per field, 14 mm square field size, 9 points per wafer, 150 mm wafer, reticle change every 10 wafers, 100 reticle changes. This gives 81,000 data points per axis and 9000 total fields.

OVS was used to estimate the type of control necessary for individual component errors in order to reach overlay design rules commensurate with 0.5 µm minimum feature size. It is generally agreed that overlay must be in the 0.1 to 0.15 µm, 3 sigma range for 0.5 micron lithography.

First considered is a stepper using global alignment. The input error components used for the simulation are listed in Table 4. The first column in Table 5 lists the good fields percentage and the mean plus 3 sigma for the distribution of all errors for this stepper. The same number of simulation runs and locations as in the previous example were used. 99.85% of all fields meet the 0.1 µm requirement. Note the very high precision necessary for all the component errors, better than is currently available on any stepper system. In particular translation error control to 0.021 µm, 3 sigma, is about 3 to 4 times better than any reported and stage error control to 0.03 µm, 3 sigma, is about 2 to 3 times better. Mask error control of 0.045 µm, 3 sigma, is only possible for reduction reticles. Substantial improvements are necessary in order to reach 0.1 µm single stepper overlay if global alignment is used.

Table 4 - Characteristic Component Overlay Errors; 0.1 µm Stepper

Interfield Errors	Units	Global Stepper	F X F Stepper
Translation offset sigma	µm	0.005	0.015
Translation sigma around offset	µm	0.005	0.015
Symmetrical expansion offset sigma	ppm	0.25	0
Asymmetrical expansion offset sigma	ppm	0.25	0
Expansion sigma around offset	ppm	0.25	0
Wafer rotation offset sigma	ppm	0.25	0
Orthogonality offset sigma	ppm	0.25	0
Rotation sigma around offset	ppm	0.25	0
Stage precision sigma	µm	0.01	0
Intrafield Errors			
Magnification sigma around offset	ppm	1.0	2.0
Reticle rotation sigma around offset	ppm	1.0	2.0
Reticle stacking error sigma (@ 1/5X)	µm	0.015	0.015

Table 5 - Overlay Simulation Results; 0.5 Micron Stepper
Global Alignment

Overlay Design Rule (micron)	Single Stepper % Good Fields Mean	Std. Dev.	Matched Steppers 0.05 um Distortion Mean	Std. Dev.
.025	1.01	2.95	0.00	0.00
.050	78.08	14.04	24.54	17.91
.075	97.33	4.48	65.15	20.86
.100	99.85	0.45	91.56	9.05
.125	100.00	0.00	98.86	1.85
.150	100.00	0.00	99.79	0.57
.175	100.00	0.00	100.00	0.00
$\bar{X}$ + 3 sigma calculated for all data (micron)	0.092		0.140	

OVS was then run for the 0.1 micron system to simulate matching between two systems which have a relative maximum intrafield error of 0.05 micron in both the x and y axes. Figure 8 shows the good fields percentage versus the overlay design rule for a single machine and matched steppers.

Note from the data listed in Table 5 that $\bar{X}$ + 3 sigma for the matched steppers is almost exactly the relative distortion error of 0.05 µm plus the $\bar{X}$ + 3 sigma determined for the single stepper (0.092 + 0.05 = 0.142 µm, as opposed to the simulation result of 0.140 µm). If the root sum square is taken the result is 0.105 µm. This demonstrates that it is incorrect to root sum square a systematic error such as distortion with random alignment errors to arrive at a total overlay budget, because doing so underestimates the true result, as is argued in reference (1).

Next considered is a stepper working in the field by field (FXF) alignment mode. The same field and wafer sizes, and number of points sampled are used. It is assumed that the stepper has a mechanism to adjust magnification to fit symmetrical wafer expansion perfectly, and that asymmetrical expansion is not present. In this case, errors due to wafer rotation, orthogonality, and stage error can be set to zero. The input parameters are listed in Table 5 and the simulation results are given in Table 6. Again it is found that the systematic distortion error adds directly to the result for single machines to arrive at the mean plus 3 sigma total overlay. Figure 9 shows the good fields percentage versus overlay for the FXF case.

The simulation results show that it's possible to relax the alignment error control by about three fold and the reduction error control by two if FXF alignment is used rather than global to reach 0.15 µm total overlay for mixed steppers. The level of control necessary for global alignment seems quite difficult to reach while that for FXF seems possible. The difficulty in employing FXF is usually reduced productivity. In order to achieve reasonable throughput rates, alignment acquisition times have to be quite small at each field. In addition, field alignment targets are typically much smaller than global targets because of the limited real estate available in scribe lines or in the chips themselves. Thus the signal is usually not as strong as with a large global target.

Table 6 - Overlay Simulation Results; 0.5 Micron Stepper
Field by Field Alignment

Overlay Design Rule (micron)	Single Stepper % Good Fields Mean	Std. Dev.	Matched Steppers 0.05 μm Distortion Mean	Std. Dev.
.025	0.00	0.00	0.00	0.00
.050	73.40	28.90	0.00	0.00
.075	96.10	13.25	53.20	30.46
.100	99.80	1.41	90.00	20.88
.125	100.00	0.00	98.30	1.85
.150	100.00	0.00	100.00	0.00
$\bar{X}$ + 3 sigma calculated for all data (micron)	0.095		0.144	

It is likely that field by field alignment will have to be used to achieve overlay in the 0.1 μm range, but this will require more real estate allotted to targets, more targets per field in order to allow active magnification, reticle rotation, and trapezoid control, and faster, more accurate detection and mechanical adjustment schemes than currently available. In addition, overlay simulation will be necessary to understand in detail the complex interaction of mask making, stepper, and processing variables.

Non-Gaussian Overlay Distributions

It is commonly claimed by aligner manufacturers that overlay distributions are not Gaussian, but rather have less data points in the extreme tails of the distribution than a normal distribution does. The form of a distribution is characterized in statistics as kurtosis. Kurtosis relates to the tendency for a distribution to have a sharp peak in the middle and excessive data in the tails as compared with a Gaussian or conversely to be relatively flat in the middle with little or no tails[11]. The coefficient of kurtosis, g_2, characterizes whether a distribution is leptokurtic (containing more data in the tails, $g_2 > 0$), mesokurtic (Gaussian, $g_2 = 0$), or platykurtic (less data in the tails, $g_2 < 0$). The coefficient of kurtosis, g_2, for a sample of n numbers $X_1, X_2, \ldots, X_n$ is calculated

$$g_2 = \frac{\sum (X_1 - \bar{X})^4}{ns^4} - 3 \qquad (18)$$

The coefficient of kurtosis is calculated for each simulated set of alignments (i.e., one reticle change) in OVS. It is found that there is a distribution of g_2 values for each set of error parameters. For example, in the case of the 0.1 μm overlay stepper using global alignment the distribution of g_2 along the x axis for 100 simulated runs is shown in histogram form in Figure 10a. The mean of the distribution is - 0.144, showing that indeed that the average run does have a distribution of overlay errors which is more tightly bunched than a Gaussian. However the distribution of g_2 is very wide, with a standard deviation of 0.294. 22 of 100 runs had positive g_2 values. Thus while the average lot of the 0.1 μm single stepper has a platykurtic distribution of overlay errors, almost a quarter of the runs can be expected to show leptokurtic behavior.

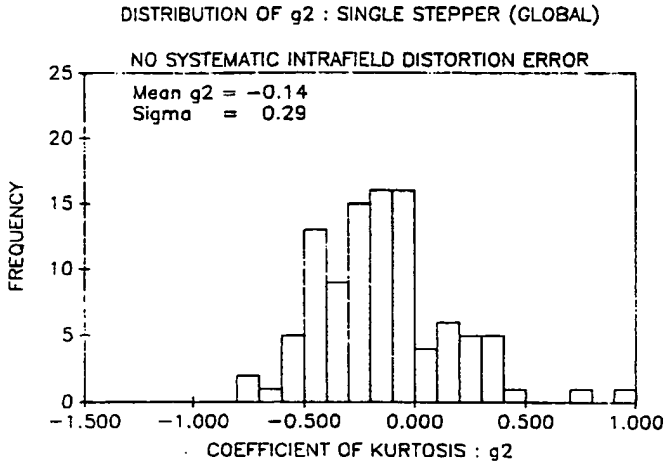


Figure 10a. g_2 histogram;global,single machine.

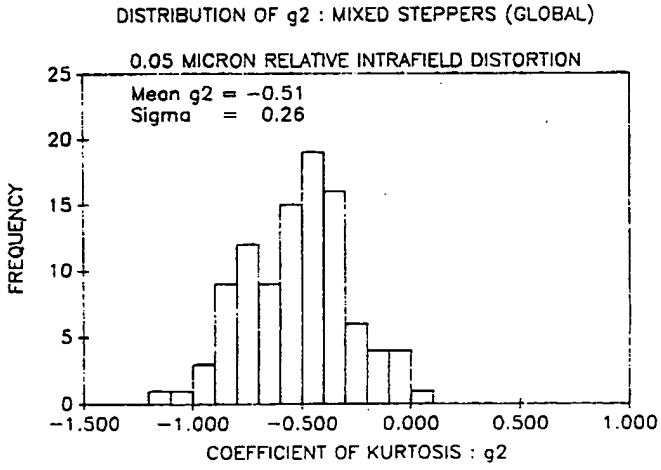


Figure 10b. g_2 histogram;global, mixed.

Table 7 - Coefficient of Kurtosis for Simulator Runs

	Single Stepper	Matched Steppers (0.05 μm distortion)
Global	-0.14 ± 0.29	-0.51 ± 0.26
FXF	-0.35 ± 0.39	-0.62 ± 0.42

For the matched stepper case the distribution of g_2 values along the x axis is shown in Figure 10b. Note that the center of the distribution is more shifted to negative g_2 values than the previous case where there was no systematic intrafield error. The mean is - 0.510 and only two of 100 runs showed positive g_2 values. The mean and one standard deviation values for g_2 for the global and FXF steppers are listed in Table 7.

These simulation results imply that overlay errors for steppers do not in general yield Gaussian distributions. However, neither do they always yield distributions which have less data in the tails than a Gaussian unless systematic errors are of about the same size as the random errors. The reason for the shift to more platykurtic distributions with increasing systematic errors is not completely clear. An intuitive argument can be given. As systematic errors grow larger with respect to the random errors, the overall distribution of errors begins to be dominated by the systematic errors. In the limit the error distribution converges with the systematic error distribution which has no randomness by definition. Thus the error distribution might be expected to be platykurtic. This is an important area for further analysis.

Acknowledgments

I would like to thank a number of people who contributed to the ideas expressed in this paper. First I would like to thank Lucien Nedzi, now with the U.C.S.F. Medical School, for his substantial contribution to the contour representation of linear overlay errors. The section on the contour representation is a condensed version of an unpublished paper[12] written with Anna Minvielle of AMD in 1984. Rory Rice of AMD has been constantly insistent that overlay error budgets take into account systematic errors and in many ways this paper is in response to his searching questions. Alan Levine of Ultratech Stepper and Colin Knight of AMD/Sematech suggested to me the idea of constructing a Monte Carlo simulator for wafer stepper overlay errors. Finally I would like to thank Bill Heavlin and Andy Brown of AMD, and Harry Levinson of Sierra Semiconductor for helpful discussions.

References

(1) R.Rice, H.Levinson,"Overlay tolerances for VLSI using wafer steppers",SPIE Vol.922,1988

(2) D.S.Perloff, "A four point electrical measurement technique for characterizing mask superposition errors on semiconductor wafers", IEEE J.Sol.St.Circ., Vol. SC-13,4,436-444,1978

(3) D.MacMillen,W.D.Ryden,"Analysis of image field placement deviations of a 5X microlithographic reduction lens", SPIE Vol.334,78-89,1982

(4) W.T.Lynch,"The reduction of LSI chip costs by optimizing the alignment yields",IEDM Technical Digest,7G-J, 1977

(5) W.H. Arnold,"Image placement differences between 10:1 reduction wafer steppers and 1:1 scanning projection aligners", SPIE Vol. 394,1983

(6) J.R. Taylor, An Introduction to Error Analysis, University Science Books

(7) W.Heavlin, C. Beck, "On yield optimizing design rules", IEEE Circuits and Devices Magazine, 7-12, March 1985

(8) T.F. Hasan, S.U. Katzman, D.S. Perloff, "Automated electrical measurements of registration errors in step-and-repeat optical lithography systems", IEEE Trans. El. Dev., Vol.ED-27, No. 12, pp2304-2312, 1980

(9) C.S. Kim, W.E. Ham, "Yield-area analysis: part 2 - effects of photomask alignment errors on zero yield loci", RCA Review, Vol. 39, 565-576, 1978

(10) D.W. Peters,"The effects of an incorrect condenser lens set-up on reduction lens printing capabilities", Procee dings of the Kodak Microelectronics Interface, 1985

(11) Manual on the Presentation of Data and Control Chart Analysis, ASTM STP 15D

(12) W.H. Arnold, A.M. Minvielle,"Distribution model and contour representation of systematic registration errors in optical lithography", 1984, unpublished paper